Extending the capabilities of DNA writing tools to improve medicine

Lear, Sierra K

Extending the capabilities of DNA writing tools to improve medicine

by

Sierra Lear

DISSERTATION

Submitted in partial satisfaction of the requirements for degree of

DOCTOR OF PHILOSOPHY

in

Bioengineering

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO
AND
UNIVERSITY OF CALIFORNIA, BERKELEY

Approved:

DocuSigned by:

Shipman, Seth ______________________ Shipman, Seth

DF90F03FBB1341E...

Chair

DocuSigned by:

John Dueber ______________________ John Dueber

[illegible]4B2...

DocuSigned by:

Ken Nakamura ______________________ Ken Nakamura

83D6801FCC3C4A4...

____

____

Committee Members

Dedication

To my beloved husband and parents, my greatest role models

Acknowledgements

I would like to thank the Shipman Lab, especially my advisor Dr. Seth L. Shipman and my colleagues and friends, Rebecca Fang, Chloe Fishman, Santiago Lopez, and Dr. Christina Palka for their unwavering moral support and nurturing scientific wisdom throughout my long PhD journey. I also must thank the Gladstone Institute Intellectual Property & Legal Affairs Department for so kindly letting me learn from them. I ended up with a much greater respect for all the work that goes on behind the scenes to allow scientists like myself to focus single-mindedly on our research.

While the PhD journey is long, it was preceded by an even longer journey to be accepted into UCSF and UC Berkeley. Without the support of my parents, my friends, and the wonderful culture and scientists fostered in the labs of Dr. Stu Tobet, Dr. Dan Gustafson, Dr. Anne Robinson, and Dr. Miriam Goodman, I would not be where I am today.

Finally, I would be deeply amiss if I did not credit my husband, David. My marriage to you is, without a doubt, the most valuable takeaway from my time in graduate school. Nor would I have even accomplished the main academic takeaway from graduate school—my thesis—without your loving presence, for you were my pillar every time an experiment failed or went awry. You also taught me how to code, a skill that I did not possess prior to attending graduate school.

Contributions

Chapter 1 contains work previously published in:

- Lear, S.K. & Shipman, S.L. Molecular recording: Transcriptional data collection into the genome. *Current Opinions in Biotechnology*. **79**, 102855 (2023). doi: 10.1016/j.copbio.2022.102855

Chapter 2 contains work previously published in:

- Lear, S.K.*, Lopez, S.C.*, González-Delgado, A., Bhattarai-Kline, S., & Shipman, S.L. Temporally resolved transcriptional recording in E. coli DNA using a Retro-Cascorder. *Nature Protocols.* (2023). doi: 10.1038/s41596-023-00819-6. *Equal contribution.

Chapter 3 contains work previously published in:

- Lopez, S.C., Crawford, K.D., Lear, S.K., Bhattarai-Kline, S. & Shipman, S.L. Precise genome editing across kingdoms of life using retron-derived DNA. *Nature Chemical Biology*. **18**, 199-206 (2022). doi: 10.1038/s41589-021-00927-y

Chapter 4 contains work from a manuscript under review:

- Lear, S.K., Nunez, J.A., & Shipman, S.L. High-throughput colocalization pipeline quantifies efficacy of mitochondrial targeting signals across different protein types. *bioRxiv* (2023). doi: 10.1101/2023.04.03.535288

Abstract

Extending the capabilities of DNA writing tools to improve medicine

Sierra Kyli Lear

DNA gene editing tools, also referred to as DNA writers, have already improved human healthcare outcomes and may continue to be leveraged to further improve human health. Interesting but unsolved medical applications include using DNA writers to record cellular development, which underpins numerous diseases, and to cure mitochondrial genetic disease. Although gene editing techniques have previously been used to cure nuclear genomic diseases, delivery of DNA template and proteins into mitochondria hinders our ability to directly edit mitochondrial DNA. Here, we first show that the DNA writers Cas1 and Cas2 can be used to record the order of transcriptional events in cell populations and then develop a practical and accessible protocol allowing others to also implement the technique. Next, we show that combining Cas9 with a retron reverse transcription enables researchers to achieve precise edits, including exon-long insertions, in human cultured cells and reduce our reliance on exogenous delivery of DNA template, a barrier in mitochondrial gene editing. Finally, we develop a useful methodological pipeline that can quickly quantify the colocalization of different engineered proteins in mitochondria using immunocytochemistry, high-throughput fluorescent imaging, and automated analysis written in Python. We hope this resource may help researchers empirically engineer new DNA writers that are efficiently imported into human mitochondria. Overall, the new genetic technologies and pipelines described here may assist future scientists in engineering new gene editing approaches to cure developmental or mitochondrial disorders.

Table of Contents

List of Figures

List of Supplementary Figures

List of Tables

Chapter 1: Introduction

1.1 Extending the capabilities of DNA writers to improve medicine

Since the discovery of flexible and easily programmable DNA writing tools, such as CRISPR/Cas9, there has been an explosion in interest in using DNA writers to develop novel medical treatments[1]. However, these therapies often require further optimization of the DNA writers on which they depend, typically by combining the strengths of the DNA writer with additional proteins with complementary functions or engineering the DNA writers to diversify their effect or localization. In this thesis, I used both strategies to build biotechnology to leverage DNA writing for applications related to regenerative medicine and therapeutic mitochondrial genome editing.

One key goal of regenerative medicine is to replace damaged tissues by growing replacement parts *in vitro*. However, growing a specific cell type in the lab requires an intimate understanding of its cellular development, specifically the transcriptional signals and cues that guide an undifferentiated stem cell towards a desired and differentiated cell type. In **Chapter 2**, I describe an accessible protocol to record the timing of multiple transcriptional signals in a population of *E. coli*, which is one key step in developing a DNA writing-based technology that can reveal the necessary transcriptional steps underlying cellular development.

In addition, DNA writing tools allow scientists to directly cure genetic diseases by editing a patient's diseased DNA. Although researchers have successfully used DNA writing technology to ameliorate some diseases, as illustrated by the many successful clinical trials that use genome editing to treat blood-related disorders[2], more improvements are required in order to broaden the number of diseases that genome editing can cure. Thus, I highlight technologies I developed to help enable mitochondrial DNA (mtDNA) editing by: 1) more precise genome editing through the *in vivo* production of DNA donor templates within mammalian cells (**Chapter 3**), and 2) screening how to more specifically and strongly import DNA writers into mammalian mitochondria (**Chapter 4**).

1.2 Using molecular recording to understand regenerative medicine

The journey from the zygote to fully differentiated cell relies on a series of temporally choreographed transcriptional events. Understanding the precise order of events that yields one cell type versus another is critical to advance our knowledge of basic developmental biology. Moreover, this knowledge has practical ramifications for regenerative medicine, as the sequence of events that unfolds in a developing cell may be mimicked *in vitro* to produce replacement parts for degenerative diseases. Unfortunately, conventional transcription assays like RNA-seq and in situ hybridization are not well suited to understand long and complex processes like development because they require destruction of cells in the midst of the event for analysis. To reassemble the resulting transcriptional snapshots into a continuous process requires analytical assumptions that are not always true[3].

An emerging set of technologies aims to solve this problem by recording biological data into a molecular record (DNA, RNA, or protein) that remains inside the cell during the process of interest. Since the data collection is non-destructive, the overarching biological process plays out from beginning to end, after which the data can be collected by imaging or sequencing. Much work remains to be completed to realize the lofty goal of recording all biological data into molecular records. However, a suite of useful molecular parts is emerging. Over short timescales (minutes to hours), the age of individual transcripts can be encoded onto the transcripts themselves using RNA deaminases, and the order of selected transcriptional events can be recorded as fluorescent tags incorporated into elongating protein polymers[4–6]. Over longer timescales (hours to days), DNA is the recording medium of choice. DNA-based dynamic lineage recorders (reviewed in refs. [7,8]) use CRISPR nucleases to diversify sites in the genome to encode the relationship of cells over multiple generations. In this section, we will focus on a particular suite of molecular recording technologies: those that aim to record the *order* of transcriptional events over long timescales by writing a DNA-based molecular record.

The potential impact of transcriptional molecular recording in DNA is illustrated by Cre and FLP recombinase-based reporters—ubiquitous tools within developmental biology—that can be considered simple transcriptional molecular recorders. These systems link a transcriptional event to the expression of a recombinase that makes a permanent genomic modification to a cell. Thus, the occurrence of an event is stably recorded in DNA and can be read-out at a later point in time, often by using the genomic modification to turn on a fluorescent protein. These reporter lines have been invaluable to identify specific cell populations that rely on a given transcription factor to define their cell fate[9,10].

Yet, despite the clear value of these early recorders, they are limited to recording only as many independent events as the number of fluorescent proteins that can be resolved simultaneously. Moreover, they only encode the *occurrence* of an event, but not *when* it happened. By omitting the fluorescent readout and focusing instead on the mark made to the genome as the data itself, the number of distinct signals recorded can be further extended. Moreover, if these marks to a genome are organized sequentially, event order can also be determined to yield a richer understanding of complex cellular processes. We will focus on three molecular strategies for such recordings using distinct molecular components: (1) recombinases; (2) reverse transcriptases (RTs) and CRISPR integrases; and (3) RTs and CRISPR nucleases

(**Fig. 1.1**). Despite the differences in encoding, we will argue in the concluding section that a common data structure is emerging across all strategies.

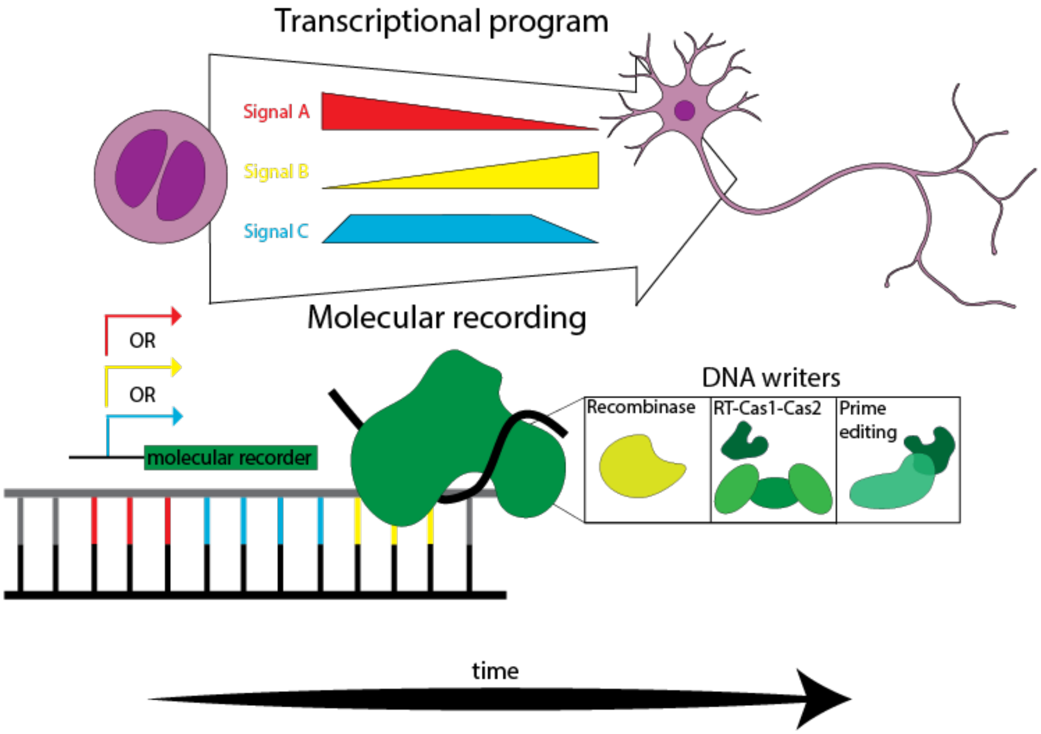


Figure 1.1. DNA-based molecular recording strategies. Over the course of cellular development, different transcriptional signal will be turned on and off in a cell. One method to record the chronology of these transcriptional events is by using a molecular recorder, where the expression of different transcriptional signals culminates in the specific edit within the genome of a cell. The most common strategies to create transcriptional genomic records rely on three different DNA writers: recombinases, RT-Cas1-Cas2, and prime editors.

1.2.1 Recombinases

Recombinase-based molecular recorders rely on the ability of DNA recombinases to flip or delete a DNA sequence surrounded by two recognition sites as a genetic mark or output. Additional layers of complexity can be added by combining multiple recombinases, promoters, and terminators to create circuits capable of responding with different genetic outputs depending on the number, order, and mixture of distinct inputs[11–15] (**Figure 1.2a**). This approach enables the development of bacterial sentinel cells that can be used as biosensors to analyze human samples, for instance detecting too much glucose in urine, a sign of diabetes[16]. Furthermore, multiple recombinase circuits have been validated in mammalian cells[17,18,15].

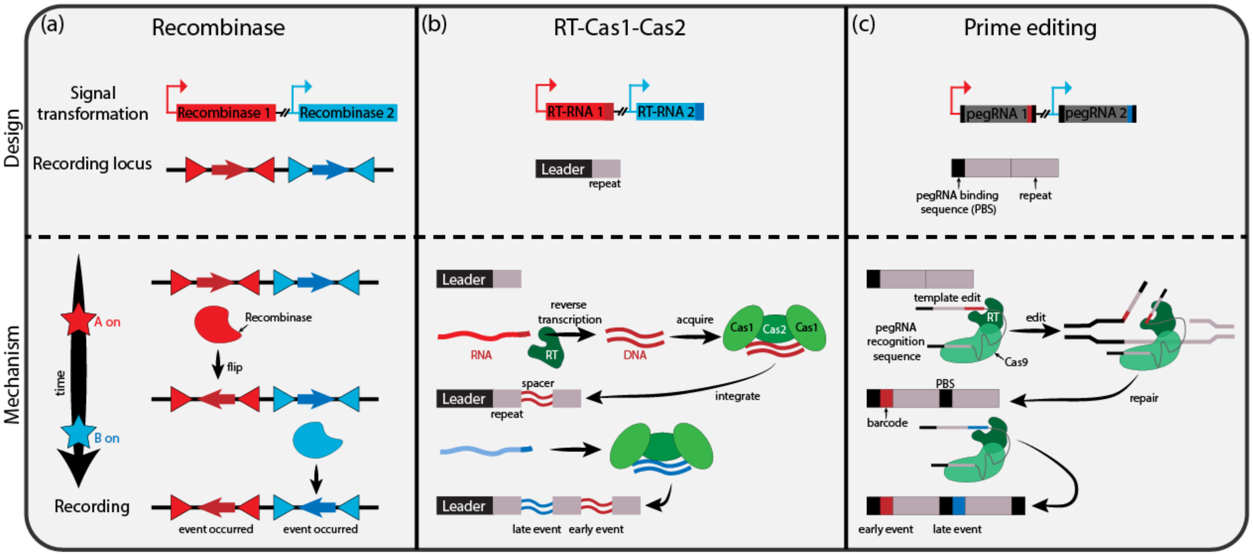


Figure 1.2. Three different DNA writers enable transcriptional recording. Transcriptional molecular recorders rely on the activity of a DNA writer to record different transcriptional events. **(a)** Site-specific recombinases are expressed from a promoter of interest and modify DNA between two recognition sites. Each edit from an orthogonal recombinase indicates a transcriptional event, but the order of events cannot be reconstructed. **(b)** Cellular or barcoded RNAs are expressed from a promoter of interest and reverse-transcribed into DNA pre-spacers that can be acquired and integrated into a CRISPR array via Cas1-Cas2. Since each spacer is always integrated next to the leader sequence, the order of events can be inferred. **(c)** Barcoded pegRNAs are expressed from promoters of interest. The pegRNA directs a prime editor to the pegRNA binding sequence (PBS) in a pre-engineered array, where it incorporates a barcode corresponding to a transcriptional event, obfuscates the previous PBS, and adds a new PBS to which the next prime editor can bind. This design allows the order of events to be inferred.

Early recombinase recorders were constructed to record whether a transcriptional event occurs rather than an event's relevant analog characteristics, such as duration or intensity. To overcome this limitation, several recombinase-based recorders were developed to more closely mimic an analog recorder, by increasing the total number of recording cells or the number of recording plasmids per cell[19,20]. While the presence of a genomic mark in a single bacterium or plasmid can only encode the occurrence of a signal, the number or percentage of such events in a larger population of loci can reflect the duration or intensity of a signal. A recurring critique of recombinase-based recorders has been that the number of independently recorded events is limited to the number of orthogonal recombinases[21]. However, a recent tweak on the approach used catalytically inactive Cas9 to direct a single recombinase to integrate distinct sequences into an expanding genomic array, depending on the gRNA expressed[22]. This design enabled duration and intensity recordings of multiple transcriptional events simultaneously, but since recombinases do not inherently have a writing direction, additional molecular components will be required to reliably reconstruct event order.

1.2.2 RTs and CRISPR Integrases

An alternative DNA writer with inherent directionality uses the CRISPR integrases Cas1 and Cas2. These integrases act as part of a bacterial immune system that acquires phage DNA sequences into a genomic repository as an immunological memory of that phage. This genomic repository, called the CRISPR array, consists of a short leader sequence followed by a series of unique fragments of DNA from foreign invaders, called spacers, separated from each other by repetitive DNA sequences called repeats. During acquisition, a complex of Cas1 and Cas2 can integrate spacers into the CRISPR array next to the leader sequence[23,24]. Since the Cas1-Cas2 complex always inserts its newly caught spacer next to the leader, the recorded spacers are also captured in chronological order, where the oldest events are those furthest away from the leader. Early technological work in this area showed that the Cas1-Cas2 acquisition system can be hijacked to store electroporated synthetic oligonucleotides as spacers, even if these spacers contained information unrelated to the immune system—such as data encoding a video[25,26].

1.2.2.1 Transcriptional event recording via RT-Cas1-Cas2

Unfortunately, most natural CRISPR systems integrate DNA, rather than the RNA that would be required for a transcriptional recorder. One workaround to this problem is to use a biological signal to modulate the copy number of a plasmid containing pre-spacer sequences[27]. This approach was later modified by replacing the biological signal with an electrical signal to encode data in bacteria for industries in need of secure data[28]. Alternatively, other recent strategies instead convert RNA into DNA using an RT (**Figure 1.2b**).

Components of a CRISPR system from *M. mediterranea* (MMB-1), including a natural RT-Cas1 fusion, were shown to acquire spacers derived from donor RNA in MMB-1[29]. RT-Cas1 has now been translated into a recording technology[30–32]. Using an *F. saccharivorans* RT (FsRT)-Cas1 fusion which can promiscuously reverse-transcribe most RNA transcripts into potential spacers, this technology captures a diverse set of RNA-derived spacers into a population of CRISPR arrays and thus extends the global transcriptomic snapshot provided by RNA-seq to a global transcriptomic *history*. Proof-of-concept work demonstrated differentiable transcriptomic histories of bacteria that were or were not transiently exposed to a herbicide[30].

In addition to using Cas1-Cas2-based transcriptional recorders to develop a global transcriptomic history, we showed that they can also be used to infer the ordering of two specific promoters of interest using a method called Retro-Cascorder[33]. This approach replaces the promiscuous FsRT with a retron RT, which only reverse transcribes its own non-coding RNA (ncRNA). Plasmids were engineered to express two inducible promoters of interest, each linked to a different barcoded retron ncRNA. The ordering of the barcoded spacers in individual genomic arrays can be used to determine the order in which different promoters were previously induced. In **Chapter 2**, I will describe a protocol to enable other researchers to perform recording using a Retro-Cascorder in further depth.

1.2.2.2 Porting Cas1-Cas2-based molecular recorders to mammalian cells

Although certain elements relevant to transcriptional recording, such as the retron RT, are able to reverse-transcribe RNA and even mediate editing in eukaryotic cells, including yeast and human cells[34–37], Cas1 and Cas2 have so far not been shown to be functional in eukaryotic cells. This limitation to porting Cas1-Cas2-based recording systems into mammalian cells, which may rely on host factors like *E. coli* immune host factor[38,39], must still be solved to increase the impact of this technology.

1.2.3 RTs and CRISPR Nucleases

A CRISPR component that has had no trouble being ported into eukaryotic cells is Cas9. Previous event recorders have already capitalized on Cas9 or another CRISPR nuclease, Cas12a, to link specific biological stimuli to edits in DNA, but these technologies do not capture the chronological order of the events[40–44] (see also refs. [45,46] for in-depth reviews of how CRISPR nucleases have been used as event recorders). However, by combining Cas9's flexible operation in multiple cell types with the architectural strengths of a CRISPR array and an RT to convert RNA signals into DNA, new recording systems based on prime editing[47] have been created in mammalian cells (**Figure 1.2c**).

A prime editor (PE) consists of a Cas9 nickase fused to an RT. The Cas9 nicks at a specific site encoded on its accompanying prime editing guide RNA (pegRNA). Afterwards, the RT reverse primes off the exposed genomic cut site to reverse-transcribe an edit-encoding extension on the pegRNA and create a precise edit in the target site[47]. To build a molecular recorder, the edit-encoding extension of the pegRNA can be engineered to include a sequence that encodes a barcode followed by a pegRNA-binding sequence that a future pegRNA needs to mediate the next edit. As a result, a series of barcodes, all encoded by unique pegRNAs expressed under different promoters or signals of interest, can be added in an ordered fashion to a chosen genomic locus, much like the architecture of a CRISPR array.

Three recent papers have used this prime editing-based strategy to develop a PE-based molecular recorder within mammalian cells that can accurately encode information[48] and measure the strength and intensity of different activated signaling pathways, such as Wnt[49]. Additionally, after sequentially transfecting cells with plasmids expressing pegRNA, PE-based recorders accurately captured the order of the nucleic acid delivery[48,50].

Although the use of prime editing overcomes the limitation of porting Cas1-Cas2-based molecular recorders to eukaryotic cells, current recorders using PEs must either pre-build an array of a defined length[48] or use multiple pegRNA-binding sequences that switch back and forth[50], creating arrays that are less open-ended than the CRISPR arrays that were the inspiration. Moreover, inefficiencies in barcode insertion still need to be overcome to capture ordered biological information within living cells with these systems.

1.2.4 Outlook & outstanding challenges

The quest to build an ideal transcriptional molecular recorder has spawned numerous molecular incarnations. Each DNA writer has its own unique strengths and weaknesses. Recombinases are comparatively efficient, with a better chance of recording rare or transient events, but are more difficult to scale and lack directionality. Cas1-Cas2-based approaches are nearly agnostic to the number of barcodes and use an open-ended array, which bodes well for scalability and recording duration, but are currently limited to bacteria. Prime editing-based approaches are multiplexable and deployable in eukaryotes, but have a less open-ended data array, and need improvements in efficiency to record event timing.

Nonetheless, the most modern and promising transcriptional recordings share a common recording infrastructure to store transcriptional information. This data structure can be described as a CRISPR-like array which continually expands in a single direction by adding barcodes corresponding to a transcriptional signal. Such an infrastructure is highly multiplexable, stores high-density data, and conveys clear event ordering.

However, such a data structure also carries some inherent limitations. First, arrays containing multiple signals of interest, and thus clear event ordering information, are very rare. Since the probability of integrating multiple spacers is multiplicative, the chances of creating an array with two or three signals of interest recorded is exponentially rarer than an array with one signal of interest. This issue is further compounded by the low integration efficiencies of current DNA writers like Cas1-Cas2 and PE. These multi-spacer arrays may still be captured by increasing sequencing depth or using methods like SENECA to specifically enrich and amplify expanded CRISPR arrays[30,31]. Nonetheless, the rarity does limit the ability of any expanding array-based recorder to approach single-cell resolution. One solution is to optimize recording components to increase efficiency so that arrays with multiple informative events are more frequent, but there are likely inherent limits on that optimization (e.g., the refractory time it takes to repair a genomic array after barcode integration).

Second, deconvolving signal intensity and duration is extremely difficult. Analyses of molecular recordings that use expanding arrays are influenced by both characteristics, since each of them result in an increased number of integrated spacers[33,48]. Instead, more nuanced methods or multi-dimension readouts are necessary to describe a signal that occurs earlier but at a lower intensity compared to a signal that occurs later but at a higher intensity. In future incarnations, including an additional timestamp within the spacer sequence could help differentiate signal duration from intensity.

Finally, developing an expanding array-based recorder will almost always require some amount of pre-engineering, both to insert signals into a genomic recording locus and to drive the machinery that performs the DNA writing. The recording may also place a burden on its host that could perturb the transcriptomic changes that researchers aim to capture. This limitation may prevent transcriptional recording technologies from being used directly in the cells of interest. Rather, engineered sentinel cells may be a more tractable approach to understanding certain aspects of human health[16,51,32].

Extensive research into different molecular components has mediated impressive gains in the scalability and transferability of transcriptional recorders, and continued optimization will likely result in greater efficiency and resolution. However, given an expanding array-based data structure's inherent constraints, investigating alternative architectures for biological data storage is a promising method to further revolutionize transcriptional molecular recording. Additional progress in incorporating direct- or random-access memory to biological data storage could, for example, reveal novel ways to approach single-cell resolution and unambiguous information content. After overcoming these key barriers, transcriptional molecular recorders are poised to become a widespread and invaluable tool in understanding gene expression during complex events like development.

1.3 Engineering DNA writers to enable mitochondrial gene editing

The mitochondrial genome is a circular genome consisting of approximately 16,000 nucleotides that encode proteins critical for oxidative phosphorylation. Despite its small size when compared to the nuclear genome, mtDNA is much more likely to accrue pathogenic mutations than nuclear DNA[52]. Moreover, mtDNA mutations can lead to severe consequences and is implicated in aging and numerous neurodegenerative diseases[53,54]. In fact, around 1 in 5000 individuals suffer from a mitochondrial genetic disease[53,55].

Given the phenotypic consequences and frequency of mtDNA mutations, therapeutic mitochondrial genome editing appears to be a promising cure. However, unlike the nuclear genome, we lack efficient ways to specifically alter the mitochondrial genome. Strategies to manipulate mammalian mtDNA have traditionally been limited to altering mtDNA heteroplasmy[56] or relying on clinically-isolated patient samples as a genetic source. Recently, a mitochondrial-targeted cytidine base editor was described, which can alter mtDNA[57]. However, this base-editing is limited to C-G to T-A conversions, unlike other gene-editing techniques which can delete or add entire new sequences. Furthermore, this tool was reported to also result in deleterious off-target editing in the nuclear genome[58].

Two obstacles hindering the successful implementation of flexible mitochondrial gene editing in mammalian cells that this thesis aims to tackle are: 1) lack of non-template-based double-stranded break (DSB) repair pathways in mitochondria, and 2) inefficient import of reagents necessary for mitochondrial gene editing to mitochondria.

1.3.1 Mitochondria lack non-homologous end joining

CRISPR-based editing technologies typically rely on a nuclease that cuts the DNA at a specific site, inducing a DSB repair. In the nuclear DNA of mammalian cells, this DSB repair pathway most often recruited to fix the cut is called non-homologous end joining (NHEJ). This strategy often leads to a non-specific insertion or deletion at the cut site, often called an indel. Alternatively, in the presence of a DNA donor template with homology to either end of the broken DNA strands, homologous recombination (HR) may lead to the incorporation of the precise edit encoded on the donor into the DNA[59]. This HR-based strategy is often referred to as template-based or precise genome editing.

Although NHEJ is the most common DSB repair pathway for the mammalian nuclear genome, evidence for its existence in mammalian mitochondria is murky[60]. Instead, cut mitochondrial DNA molecules are typically degraded, a key reason why original mtDNA editing strategies focused on shifting mtDNA heteroplasmy levels by eliminating mutated mtDNA from rather than correcting the mutation itself[56,61]. However, mammalian mitochondria can use HR, meaning that precise editing may be viable way to flexibly correct diseased mtDNA sequences[62]. As a first step towards mitochondrial gene editing, I developed both plasmid- and RNA-based strategies to enable *in vivo* production of DNA donor templates for precise editing of the nuclear genome in **Chapter 3**.

1.3.2 mtDNA editing reagents are not specifically imported into mammalian mitochondria

Despite the success of precise DNA editing strategies in the nuclear genome, these methods have not yet been reliably translated to the mitochondrial genome. Rather, researchers struggle to successfully import both nucleic acids, such as guide RNAs, and proteins into mammalian mitochondria. While the localization of nucleic acids in mitochondria is considered the key obstacle[63], localizing some non-natural proteins, especially those that are hydrophobic, still remains a challenge[64,65]. Antón et al. reports inefficient localization of specific CRISPR nucleases in mammalian mitochondria[66]. Additionally, non-specific localization of mitochondrial gene editing technology to the nuclei can lead to dangerous off-target effects[58]. In **Chapter 4**, I focus on screening different engineered proteins to quantify which proteins relevant to template-based editing best localize in mammalian mitochondria.

1.4 References

1. Pickar-Oliver, A. & Gersbach, C. A. The next generation of CRISPR–Cas technologies and applications. *Nat. Rev. Mol. Cell Biol.* **20**, 490–507 (2019).
2. Urnov, F. D. The Cas9 Hammer—and Sickle: A Challenge for Genome Editors. *CRISPR J.* **4**, 6–13 (2021).
3. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* **18**, 723–732 (2021).
4. Rodriques, S. G. *et al.* RNA timestamps identify the age of single molecules in RNA sequencing. *Nat. Biotechnol.* **39**, 320–325 (2021).
5. Linghu, C. *et al.* Recording of cellular physiological histories along optically readable self-assembling protein chains. 2021.10.13.464006 Preprint at https://doi.org/10.1101/2021.10.13.464006 (2021).
6. Lin, D. *et al.* Time-tagged ticker tapes for intracellular recordings. 2021.10.13.463862 Preprint at https://doi.org/10.1101/2021.10.13.463862 (2021).
7. Woodworth, M. B., Girskis, K. M. & Walsh, C. A. Building a lineage from single cells: genetic techniques for cell lineage tracking. *Nat. Rev. Genet.* **18**, 230–244 (2017).
8. McKenna, A. & Gagnon, J. A. Recording development with single cell dynamic lineage tracing. *Development* **146**, dev169730 (2019).
9. Greig, L. C., Woodworth, M. B., Galazo, M. J., Padmanabhan, H. & Macklis, J. D. Molecular logic of neocortical projection neuron specification, development and diversity. *Nat. Rev. Neurosci.* **14**, 755–769 (2013).
10. Franco, S. J. *et al.* Fate-Restricted Neural Progenitors in the Mammalian Cerebral Cortex. *Science* **337**, 746–749 (2012).

11. Siuti, P., Yazbek, J. & Lu, T. K. Synthetic circuits integrating logic and memory in living cells. *Nat. Biotechnol.* **31**, 448–452 (2013).

12. Bonnet, J., Yin, P., Ortiz, M. E., Subsoontorn, P. & Endy, D. Amplifying Genetic Logic Gates. *Science* **340**, 599–603 (2013).

13. Yang, L. *et al.* Permanent genetic memory with >1-byte capacity. *Nat. Methods* **11**, 1261–1266 (2014).

14. Roquet, N., Soleimany, A. P., Ferris, A. C., Aaronson, S. & Lu, T. K. Synthetic recombinase-based state machines in living cells. *Science* **353**, aad8559 (2016).

15. Weinberg, B. H. *et al.* Large-scale design of robust genetic circuits with multiple inputs and outputs for mammalian cells. *Nat. Biotechnol.* **35**, 453–462 (2017).

16. Courbet, A., Endy, D., Renard, E., Molina, F. & Bonnet, J. Detection of pathological biomarkers in human clinical samples via amplifying genetic switches and logic gates. *Sci. Transl. Med.* **7**, 289ra83-289ra83 (2015).

17. Prochazka, L., Angelici, B., Haefliger, B. & Benenson, Y. Highly modular bow-tie gene circuits with programmable dynamic behaviour. *Nat. Commun.* **5**, 4729 (2014).

18. Lapique, N. & Benenson, Y. Digital switching in a biosensor circuit via programmable timing of gene availability. *Nat. Chem. Biol.* **10**, 1020–1027 (2014).

19. Farzadfard, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272 (2014).

20. Hsiao, V., Hori, Y., Rothermund, P. W. & Murray, M. M. A population-based temporal logic gate for timing and recording chemical events. *Mol. Syst. Biol.* **12**, 869 (2016).

21. Yehl, K. & Lu, T. Scaling computation and memory in living cells. *Curr. Opin. Biomed. Eng.* **4**, 143–151 (2017).

22. Shur, A. & Murray, R. M. *Proof of concept continuous event logging in living cells*. http://biorxiv.org/lookup/doi/10.1101/225151 (2017) doi:10.1101/225151.

23. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40**, 5569–5576 (2012).

24. Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).

25. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).

26. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345–349 (2017).

27. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science* **358**, 1457–1461 (2017).

28. Yim, S. S. *et al.* Robust direct digital-to-biological data storage in living cells. *Nat. Chem. Biol.* **17**, 246–253 (2021).

29. Silas, S. *et al.* Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase–Cas1 fusion protein. *Science* **351**, aad4234 (2016).

30. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).

31. Tanna, T., Schmidt, F., Cherepkova, M. Y., Okoniewski, M. & Platt, R. J. Recording transcriptional histories using Record-seq. *Nat. Protoc.* 1–27 (2020) doi:10.1038/s41596-019-0253-4.

32. Schmidt, F. *et al.* Noninvasive assessment of gut function using transcriptional recording sentinel cells. *Science* **376**, eabm6038 (2022).

33. Bhattarai-Kline, S. *et al.* Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature* (2022) doi:10.1038/s41586-022-04994-6.

34. Sharon, E. *et al.* Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* **175**, 544-557.e16 (2018).

35. Kong, X. *et al.* Precise genome editing without exogenous donor DNA via retron editing system in human cells. *Protein Cell* **12**, 899–902 (2021).

36. Lopez, S. C., Crawford, K. D., Lear, S. K., Bhattarai-Kline, S. & Shipman, S. L. Precise genome editing across kingdoms of life using retron-derived DNA. *Nat. Chem. Biol.* **18**, 199–206 (2022).

37. Zhao, B., Chen, S.-A. A., Lee, J. & Fraser, H. B. Bacterial Retrons Enable Precise Gene Editing in Human Cells. *CRISPR J.* **5**, 31–39 (2022).

38. Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol. Cell* **62**, 824–833 (2016).

39. Yoganand, K. N. R., Sivathanu, R., Nimkar, S. & Anand, B. Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Nucleic Acids Res.* **45**, 367–381 (2017).

40. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science* **353**, aag0511 (2016).

41. Frieda, K. L. *et al.* Synthetic recording and in situ readout of lineage information in single cells. *Nature* **541**, 107–111 (2017).

42. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360**, eaap8992 (2018).

43. Loveless, T. B. *et al.* Lineage tracing and analog recording in mammalian cells by single-site DNA writing. *Nat. Chem. Biol.* **17**, 739–747 (2021).

44. Kempton, H. R., Love, K. S., Guo, L. Y. & Qi, L. S. Scalable biological signal recording in mammalian cells using Cas12a base editors. *Nat. Chem. Biol.* 1–9 (2022) doi:10.1038/s41589-022-01034-2.

45. Schmidt, F. & Platt, R. J. Applications of CRISPR-Cas for synthetic biology and genetic recording. *Curr. Opin. Syst. Biol.* **5**, 9–15 (2017).

46. Ishiguro, S., Mori, H. & Yachie, N. DNA event recorders send past information of cells to the time of observation. *Curr. Opin. Chem. Biol.* **52**, 54–62 (2019).

47. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* (2019) doi:10.1038/s41586-019-1711-4.

48. Choi, J. *et al.* A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* **608**, 98–107 (2022).

49. Chen, W. *et al.* Multiplex genomic recording of enhancer and signal transduction activity in mammalian cells. 2021.11.05.467434 Preprint at https://doi.org/10.1101/2021.11.05.467434 (2021).

50. Loveless, T. B. *et al. Molecular recording of sequential cellular events into DNA*. http://biorxiv.org/lookup/doi/10.1101/2021.11.05.467507 (2021) doi:10.1101/2021.11.05.467507.

51. Munck, C., Sheth, R. U., Freedberg, D. E. & Wang, H. H. Recording mobile DNA in the gut microbiota using an Escherichia coli CRISPR-Cas spacer acquisition platform. *Nat. Commun.* **11**, 95 (2020).

52. Lawless, C., Greaves, L., Reeve, A. K., Turnbull, D. M. & Vincent, A. E. The rise and rise of mitochondrial DNA mutations. *Open Biol.* **10**, 200061 (2020).

53. Taylor, R. W. & Turnbull, D. M. Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.* **6**, 389 (2005).

54. Suomalainen, A. & Battersby, B. J. Mitochondrial diseases: the contribution of organelle stress responses to pathology. *Nat. Rev. Mol. Cell Biol.* **19**, 77–92 (2018).

55. Ng, Y. S. & Turnbull, D. M. Mitochondrial disease: genetics and management. *J. Neurol.* **263**, 179–191 (2016).

56. Hashimoto, M. *et al.* MitoTALEN: A General Approach to Reduce Mutant mtDNA Loads and Restore Oxidative Phosphorylation Function in Mitochondrial Diseases. *Mol. Ther.* **23**, 1592–1599 (2015).

57. Mok, B. Y. *et al.* A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* (2020) doi:10.1038/s41586-020-2477-4.

58. Wei, Y. *et al.* Mitochondrial base editor DdCBE causes substantial DNA off-target editing in nuclear genome of embryos. *Cell Discov.* **8**, 1–4 (2022).

59. Yeh, C. D., Richardson, C. D. & Corn, J. E. Advances in genome editing through control of DNA repair pathways. *Nat. Cell Biol.* **21**, 1468–1478 (2019).

60. Kazak, L., Reyes, A. & Holt, I. J. Minimizing the damage: repair pathways keep mitochondrial DNA intact. *Nat. Rev. Mol. Cell Biol.* **13**, 659–671 (2012).

61. Peeva, V. *et al.* Linear mitochondrial DNA is rapidly degraded by components of the replication machinery. *Nat. Commun.* **9**, 1727 (2018).

62. Dahal, S., Dubey, S. & Raghavan, S. C. Homologous recombination-mediated repair of DNA double-strand breaks operates in mammalian mitochondria. *Cell. Mol. Life Sci.* **75**, 1641–1655 (2018).

63. Gammage, P. A., Moraes, C. T. & Minczuk, M. Mitochondrial Genome Engineering: The Revolution May Not Be CRISPR-Ized. *Trends Genet.* **34**, 101–110 (2018).

64. Claros, M. G. *et al.* Limitations to in vivo Import of Hydrophobic Proteins into Yeast Mitochondria. *Eur. J. Biochem.* **228**, 762–771 (1995).

65. Daley, D. O., Clifton, R. & Whelan, J. Intracellular gene transfer: Reduced hydrophobicity facilitates gene transfer for subunit 2 of cytochrome c oxidase. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 10510–10515 (2002).

66. Antón, Z. *et al.* Mitochondrial import, health and mtDNA copy number variability seen when using type II and type V CRISPR effectors. *J. Cell Sci.* **133**, jcs248468 (2020).

Chapter 2: Retro-Cascorder protocol for temporally resolved transcriptional recordings

2.1 Introduction

Cells often react to internal and external stimuli through a change in gene expression. These changes can range from simple and isolated like a single gene response to a chemical or metabolic stimulus, or complex like a multi-gene transcriptional cascade during cell differentiation. Biologists have long sought insight into cells and their environments by analyzing gene expression. This analysis requires measuring the abundance of specific RNA transcripts, usually achieved by disrupting cell membranes to physically collect RNA that can be quantified. But destroying cells for analysis has an unfortunate ramification: the same cell or cell lineage cannot be analyzed at multiple points over time, and the cell must be harvested while the event is ongoing and the RNA remains. Therefore, experimenters cannot easily collect temporal data on a gene expression cascade or reconstruct a stimulus-driven event that occurred in the cell's past.

Molecular recorders are an alternative to destructive analyses of gene expression. These molecular technologies continuously record biological activity over time and store that data in DNA. By linking a biological event, like the expression of a gene, to a permanent genomic modification within the same cell, molecular recorders enable data collection throughout an entire biological process. Different molecular recorders vary in the mechanism of data recording, using recombinases[1–8], nucleases[9–11], prime editors[12–14], or integrases[15–18] (for more information on DNA-based molecular recorders, refer to refs. [19,20]). Here we focus on the Retro-Cascorder[21], which uses *Escherichia coli* Cas1-Cas2 integrases to write events, leveraging the natural directionality of these integrases to encode events in the order that they occur.

2.1.1 Mechanism of Cas1-Cas2 integration of DNA sequences

Cas1-Cas2 integrases are an essential component of the CRISPR bacterial immune system. Under phage invasion, these integrases capture small fragments of phage DNA and integrate them into a genomic repository, called the CRISPR array, where they serve as an immunological memory of the phage. The CRISPR array consists of a leader sequence followed by a series of repetitive elements, or repeats, which are separated from each other by the fragments of phage DNA, called spacers. The Cas1-Cas2 integrases are both repetitive, inserting new phage spacers without deleting old spacers, and directional, always integrating new spacers directly adjacent to the leader sequence in the CRISPR array[22,23]. Since the insertion of the spacer always occurs adjacent to the leader sequence, the chronology of spacer integration within an array can be inferred because the spacers progress from oldest to newest as they approach the leader. This property has previously enabled molecular recorders using type I-E Cas1-Cas2 from *E. coli* to decipher the ordering of multiple rounds of exogenously delivered pre-spacers in bacteria[15,24].

2.1.2 Recording RNA-derived spacers using Cas1-Cas2

One limitation of early Cas1-Cas2-based molecular recorders is that Cas1-Cas2 have only been found to integrate DNA, not the RNA that would be required to record transcriptional events. This limitation can be addressed by adding additional components to the recording system. For instance, a reverse transcriptase (RT) can reverse-transcribe RNA into DNA that Cas1-Cas2 can integrate. A natural RT-Cas1 fusion was found in *Marinomonas mediterranea* (MMB-1) that mediates the integration of RNA-derived spacers via a reverse transcribed intermediate[25]. Another RT-Cas1 fusion was found in *Fusicatenibacter saccharivorans* and used as the basis of a recording technology in *E. coli*, enabling a global transcriptome to be recorded in a cell population and retrieved at a later point in time by sequencing the CRISPR array[17,26,27]. In both cases, the chosen RT was promiscuous and enabled the acquisition of a diverse set of RNA-derived spacers. However, this distributed acquisition makes tracking the order of events nearly impossible as the likelihood of two biologically relevant spacers being acquired into the same array – which is required to recover temporal information – is exceptionally low. To track the *timing* of specific transcriptional signals using the Retro-Cascorder, we instead used RTs from a separate bacterial immune system, the retron system, whose RTs specifically reverse-transcribe from RNA sequences that contain retron recognition elements. This specificity focuses the recording on a subset of transcripts, which have a higher likelihood of being recorded into the same array.

2.1.3 Development of Retro-Cascorder

The retron RT acts specifically on a structured noncoding RNA (ncRNA), which the RT recognizes and partially reverse-transcribes into a short fragment of single-stranded DNA. We modified the retron ncRNA to generate a fragment of DNA that can be captured and integrated into the CRISPR array found within the bacterial genome by Cas1-Cas2 integrases. If this modified retron ncRNA is driven by a promoter of interest, retron-derived spacers accumulate in CRISPR arrays only when that promoter is active. These retron ncRNAs were further modified to create diversity of sequences by changing internal bases that are not involved in reverse transcription or integration. These changes create distinct barcoded retron ncRNAs that can operate within a single cell. When the expression of distinctly barcoded retron ncRNAs are driven by different promoters of interest, barcoded retron-derived spacers accumulate in the CRISPR array according to the order of the activity of different promoters. The relative order of transcriptional events can be reconstructed by sequencing the CRISPR array (ledger) at a later point in time. Thus, by marrying the unique features of Cas1-Cas2 integrases and the retron RT, Retro-Cascorder captures order information of specific biological events within living cells.

2.1.4 Applications of Retro-Cascorder

We previously demonstrated that Retro-Cascorder can successfully reconstruct the temporal relationship of induced transcriptional events in bacteria over multiple days[21]. While we decoded the expression order of multiple inducers, including anhydrotetracycline, choline chloride, and sodium salicylate, these promoters can, in principle, be replaced with other promoters of interest. Thus, this technology could enable the construction of biosensors to monitor the occurrence and order of different stimuli, such as pollutants or pathogens, in the environment. With increased engineering to improve acquisition efficiency, Retro-Cascorder may also have the resolution to track endogenous gene expression within bacteria.

2.1.5 Limitations of the protocol

Retro-Cascorder uses a common set of molecular components for molecular recording and relies only upon variable nucleotide sequences to encode different transcriptional signals. This design theoretically allows for substantially more biological events to be simultaneously recorded as compared to recombinases, whose scaling relies on a limited set of orthogonal proteins[28]. However, a reliance upon multiple plasmids to express all components from Retro-Cascorder currently limits the number of transcriptional signals that can be tested. One of these plasmids must always include high copy number expression plasmid pSBK.079 to overexpress retron RT, Cas1, and Cas2. Another necessary plasmid is a signal plasmid, whose architecture currently enables two different transcriptional events to be recorded. Although the introduction of another signal plasmid to include more transcriptional events is possible, we have found that the host burden from propagating more than two plasmids inhibits bacteria growth and prevents successful recording on relevant time scales.

The low acquisition efficiency of Cas1-Cas2, even when expressed from a high copy number expression plasmid, may also limit temporal or signal resolution because of the paucity of CRISPR arrays that will contain barcodes. While the order of transcriptional events inferred from recording analyses is typically correct when the promoters of interest express strongly and at a time scale of at least 24 hours, ordering confidence will decrease with weaker promoters and shorter experimental time scales. This issue can be mitigated by increasing the sequencing depth or the number of biological replicates. These additional measurements ensure enough CRISPR arrays containing barcodes of interest are sequenced, as long as the number of reads does not exceed the number of initial bacterial genomes harvested (see Box 1, which discusses how to pick a starting sequencing depth).

Finally, Retro-Cascorder is currently constrained to bacteria. This limitation is because type I-E Cas1-Cas2 integrase functionality has been restricted to prokaryotes. One potential explanation for its host-specific activity is its reliance on other bacterial host factors like IHF[29,30]. Further screening for additional host factors may therefore be necessary before Cas1-Cas2 acquisition is used to record transcriptional signals within eukaryotic cells. Fortunately, the other required component, the retron RT, has already been successfully used for genomic editing in multiple eukaryotic species, including yeast and mammalian cells[31–34]. Nonetheless, even if Retro-Cascorder is constrained to bacteria, this technology may still be used within sentinel cells for translational advances, including within the mammalian gut[27].

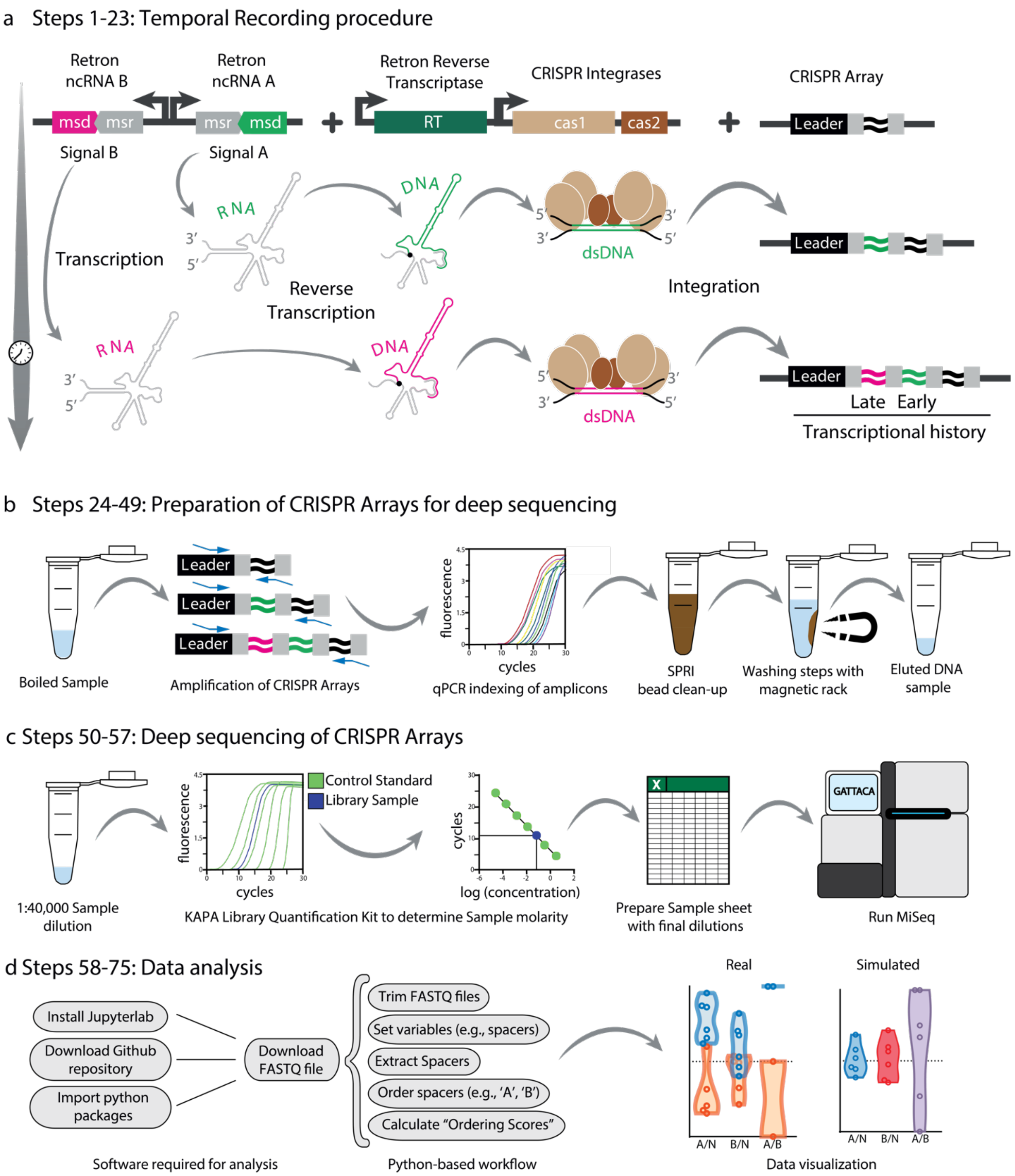
We expect that users may be interested in porting Retro-Cascorder to other bacterial species. Since we have not yet attempted to port this technology into other strains, we cannot guarantee its use in other organisms is possible. However, for those interested, the essential components that we currently know about for Retro-Cascorder include the components to create retron-derived spacers, namely a retron and RNase H, which is necessary for the retron to produce correctly sized RT-DNA[35]. Additionally, for successful acquisition of retron-derived spacers into a CRISPR array, Cas1 and Cas2, additional host factor IHF[29,30], and a CRISPR array are needed. While our protocol contains the CRISPR array in the bacterial genome, spacers can be acquired into CRISPR arrays contained on plasmids instead. However, plasmid-based acquisition efficiency is typically lower than genome-based acquisition efficiency[15], so a greater sequencing depth may be necessary to find arrays that contain information about the order of different biological events.

2.1.6 Comparison with other technologies

Adjacent technologies TRACE[16] and Record-seq[17,26] both utilize CRISPR-Cas acquisition and have also been used as biosensors to track cellular activity[36,27]. TRACE differs from Retro-Cascorder in that it uses plasmid DNA as the source of its spacers, thus making this technology most useful as a way to track or identify DNA, such as horizontal gene transfer in gut microbiota[36]. In contrast, Record-seq, similarly to Retro-Cascorder, uses an RT to convert a transcriptional signal into spacers that can be acquired by Cas1-Cas2. As discussed above, Record-seq captures a global transcriptomic profile sensitive to transient, transcriptional changes that occurred earlier in the cell's lifetime[17,27]. However, Record-seq cannot provide information about the ordering of specific transcripts of interest. Another adjacent approach called a DNA Typewriter[14] uses a prime editing strategy to modify a pre-constructed genomic locus that in principle resembles a CRISPR array. This technology generates time-ordered DNA data similar to Retro-Cascorder, but it is implemented in mammalian rather than bacterial cells. However, DNA Typewriter has thus far only been shown to resolve the relative order of transfection events, not biological signals. Given that Retro-Cascorder logs specific, pre-defined transcripts of interest, it is most useful when the desired application aims to record the transcriptional order of specifically tagged promoters of interest in bacteria.

2.2 Experimental design

The Retro-Cascorder protocol (**Fig. 2.1**) is divided into three main parts: (a) growth of bacteria containing the necessary plasmids to perform transcriptional recording, (b) preparation and deep sequencing of CRISPR arrays containing the transcriptional record, and (c) analysis of sequencing results using logical rules to infer the order of transcriptional events.



a Steps 1-23: Temporal Recording procedure

b Steps 24-49: Preparation of CRISPR Arrays for deep sequencing

c Steps 50-57: Deep sequencing of CRISPR Arrays

d Steps 58-75: Data analysis

Figure 2.1. Retro-Cascorder experimental and computational workflow. (a) Experimental procedure for Retro-Cascorder-based temporal recording. Barcoded retron-derived spacers are integrated by Cas1-Cas2 integrases into a CRISPR array only when a signal (inducer) is in the medium. The position of the spacers in the array enables reconstruction of the relative order of transcriptional events. **(b)** Preparation of CRISPR arrays for multiplexed sequencing. After acquiring genomic DNA from the samples, CRISPR arrays are selectively amplified using a

forward primer binding the leader region and a reverse primer (SPCR_MiSeq3_rev) binding an endogenous, invariant spacer always contained within the genomic CRISPR array. Amplicons are indexed using a qPCR reaction, and indexed samples are cleaned-up using SPRI beads. **(c)** Multiplexed sequencing of CRISPR arrays. Eluted samples are diluted (1:40,000) and a second qPCR reaction is used to quantify molarity of each sample using the KAPA library Quantification Kit. After preparing a sample sheet with the specifications of each sample, Retro-Cascorder libraries are deep sequenced using Illumina-MiSeq. **(d)** Computational pipeline for processing Retro-Cascorder data, beginning with installing JupyterLab Notebook, downloading Shipman's lab Github repository, and importing all python packages and dependencies required. FASTQ files from Illumina-MiSeq are processed following a python-based workflow to obtain ordering scores that may be used to generate plots to visualize both real and simulated Retro-Cascorder data. Open circles correspond to biological replicates.

2.2.1 Bacterial growth

In the first part of the protocol, BL21-AI *E. coli* containing two plasmids expressing Retro-Cascorder are grown over some specified time scale to acquire transcriptional records. Although our original publication uses a modified BL21-AI *E. coli* strain called bSLS.114 in which BL21-AI *E. coli's* endogenous retron was removed, we find that retron-based recordings still occur in the parental line. As a result, we have chosen to use the commercially available BL21-AI in the protocol due to its wider accessibility. In the case the user would like to use bSLS.114 instead, we have made it available on Addgene (catalog #191530).

The recording plasmids consist of: (a) an expression plasmid pSBK.079 constitutively expressing Eco1 RT and expressing Cas1-Cas2 under a T7 promoter, which can be induced using IPTG and l-arabinose in BL21-AI *E. coli*, and (b) a signal plasmid (e.g. pSBK.134) containing up to two promoters of interest (see Box 2, which discusses how to design signal plasmids). Each promoter expresses a modified Eco1 noncoding RNA, which can be reverse-transcribed by Eco1 RT into a barcoded DNA that Cas1-Cas2 may integrate into the BL21-AI endogenous genomic CRISPR array. Biological replicates are cultures that originated from separate, single bacterial colonies containing both plasmids.

The experimental protocol describes how to perform a transcriptional recording over 48 hours (i.e., two inducible promoters are each expressed for 24 hours). However, the protocol can be adjusted to instead record over a shorter or longer amount of time, depending on need. For longer recordings, we recommend diluting the bacterial sample into fresh LB after no more than 16 hours of growth so bacteria are kept in exponential growth conditions. Additionally, given that the strength of a promoter's signal, time scale, and sequencing depth all impact the fidelity of recordings generated using Retro-Cascorder, we recommend including a positive control in which the correct order of multiple transcriptional events is known. Our lab has previously used the signal plasmid pSBK.134 that includes two inducible promoters pTet* and pBetI, which can be turned on in either order. We recommend that users perform a 48-hour recording where they induce pTet* then pBetI for 24 hours each, or vice versa, and ensure the protocol works in their hands before moving on to perform their own experiments of interest. Although we find variability in the exact acquisition efficiency of Retro-Cascorder between different runs, this variability does not typically alter the overall trends inferred from our analysis method. However, the expression of both the signaling and expression plasmid in our strains over multiple days is burdensome on the cells and can occasionally lead to loss or corruption of the components (see Fig. 4g in our original publication[21]). As such, we recommend users always perform multiple biological replicates and to be wary when interpreting results from replicates in which there is no acquisition activity or very few informative arrays.

Following a recording experiment, the genomic records are extracted by harvesting, diluting, and lysing the bacteria. Bacterial DNA can then be stored for up to six months at -20°C until the user is ready to perform multiplexed sequencing.

2.2.2 Multiplexed sequencing

Following the recording experiment, BL21-AI CRISPR arrays are selectively amplified using PCR. Although most CRISPR arrays only contain old spacers already present within the array before the recording experiment occurred, a fraction of these arrays should also contain between one to three new spacers acquired during the recording. Some of these new spacers may be retron-derived, thus allowing event ordering to be inferred from a whole bacterial population. This first round of PCR adds Illumina adapters to the amplicons using a pool of primers with varied nucleotide lengths to diversify the samples that are eventually sequenced by the Illumina MiSeq instrument, making them compatible with downstream indexing reactions for deep sequencing. After this first round of PCR, the array-containing amplicons for each experimental condition are separately indexed, cleaned up, quantified, diluted and pooled, then finally sequenced on an Illumina MiSeq instrument. Indexing occurs using a mixture of two different primer sets; the P5 & P7 primers are added to prevent the other, longer indexing primer sets from creating too many unwanted byproducts. The MiSeq was specifically chosen to enable long enough read lengths to sequence multiple spacers in a single CRISPR array.

Given our experimental parameters (promoter strength, time resolution, and retron-derived spacer acquisition rate), we have found that sequencing each biological sample at a depth of 1 million reads allows us to reliably infer the order of transcriptional events from expanded arrays. However, in cases where the promoters are weaker or experiments are run at a short time scale, more sequencing reads may be necessary to find enough spacers to run ordering analyses.

2.2.3 Analysis

All analyses are performed using scripts written in Python 3. After quality-based read trimming, the first part of the analysis consists of extracting new spacers found in the sequenced CRISPR arrays, and storing both the new spacer sequences and the sequence of the read containing them. These reads and spacers are binned according to their characteristics, including the number of new spacers per read. Next, the order of the spacers in each newly-expanded CRISPR array is determined, as well as whether these new spacers are derived from either of the barcoded retron noncoding RNAs (referred to as "A" and "B" spacers) or not (referred to as "N" spacers, likely genome- or plasmid-derived).

For a CRISPR array to be informative for transcriptional order, it must meet three criteria: (a) the array should contain at least two new spacers, (b) at least one of the spacers should contain a barcoded retron-derived spacer, and (c) at least two spacers must have different identities. Explicitly, the number of A → B → Leader, A → N → Leader, …, etc. CRISPR arrays are counted and used for the calculation of ordering scores, described below.

Following the count of each spacer ordering possibility, we calculate three ordering scores, which describe and help us infer the ordering of transcriptional events. These scores make an assumption about the biology: that "N" spacers are acquired at a constant rate during the course of the recording experiment. Moreover, this analysis is designed to reconstruct a transcriptional history into two epochs, one early and one late. Further subdivision of the temporal signal would require a substantially more complicated analysis than is provided here.

In cases such as ours where two promoters are under study, the CRISPR arrays are analyzed for order based on three scores, each that vary between -1 to +1: (1) the A/B score, (2) the A/N score, and (3) the B/N score.

(1) A/B score. The A/B score determines both the order of and magnitude of temporal separation between the “A” and “B” transcriptional events. Positive scores suggest that transcriptional event “A” occurred before “B” and thus more “B” spacers are found in the Leader-proximal position relative to the number of “A” spacers; on the other hand, negative scores suggest the opposite, namely that event “B” occurred before “A”. The magnitude of the score represents the temporal separation between A or B, or how much the transcriptional activity between A or B overlaps in time. The more their activity overlaps in time, the closer to zero this score will be.

(2) A/N score. The A/N score determines how the timing of “A” is expressed in comparison to the constant signal “N”, for the duration of the recording experiment. It takes into account the relative frequencies of Leader-distal vs. Leader-adjacent “A”-expanded arrays: a positive score suggests that “A” was strongly expressed in the first epoch rather than the second, and conversely for a negative score.

(3) *B/N score.* The same interpretations of the A/N score applies here, except in relation to “B” rather than “A”. However, by arbitrary convention, the B/N score is reversed relative to the A/N score, calculated from the relative frequencies of Leader-adjacent vs. Leader-proximal “B”-expanded arrays: positive scores suggest that “B” occurred in the *second* epoch rather than the first, and conversely for a negative score.

Fig. 2.2 gives hypothetical examples of different transcriptional activity for two promoters across two epochs and what scores might be expected in such cases.

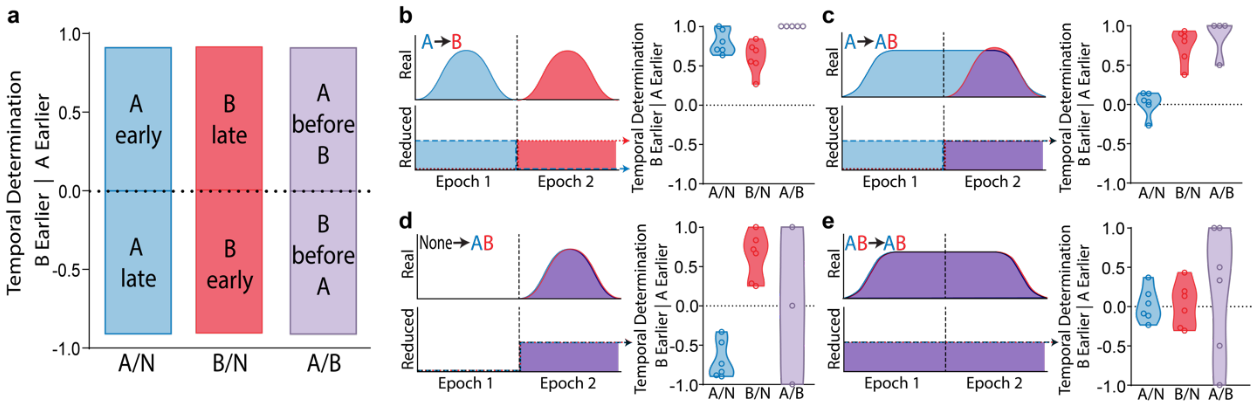


Figure 2.2. Simulated ordering score results from different transcriptional programs. Colors correspond to transcriptional events: event "A" in blue; event "B" in red; events A and B together in purple. **(a)** Key graphically illustrating how to interpret the magnitude and sign of different ordering scores. Left, ordering score A/N: a positive score suggests that event A happened, on average, before event N. The inverse is true for negative scores of A/N. Middle, ordering score B/N: a positive score suggests that event B happened, on average, after event N. Right, ordering score A/B: a positive score suggests that event A happened, on average, before event B. Panels **(b)-(e)** illustrate four transcriptional programs (left) and simulated ordering score plots (right). The transcriptional programs shown on the left of each panel show a series of "real", non-constant transcriptional signals (top), illustrated by wave-shaped curves. Below the "real" transcriptional program are a "reduced" version of these transcriptional programs, to make their inference compatible with our analysis' assumptions. On the right of panels (b)-(e) are ordering scores generated by simulating the transcriptional program described in each panel. The simulated spacer acquisition rates for "A" and "N" are equivalent to those determined experimentally in our experiments, although we have chosen to make A and B signals matched in terms of strength and leakiness. Open circles correspond to N=6 simulated biological replicates. **(b)** Transcriptional program A→B. Signal A occurs during the early epoch; signal B occurs during the late epoch (left). **(c)** Transcriptional program A→AB. Signal A is present during both the early and the late epoch; signal B occurs only during the late epoch (left). **(d)** Transcriptional program None→AB. Signals A and B occur only during the late epoch (left). **(e)** Transcriptional program AB→AB. Signals A and B occur during the early and the late epoch (left).

The command line code used in our original publication[21] is available on GitHub (https://github.com/Shipman-Lab/Spacer-Seq). We have also compiled the necessary functions with additional comments on how to perform the analysis into a Jupyter notebook, available here: https://github.com/Shipman-Lab/Spacer-Seq_Nat-Protocols/tree/main. We have made minimal changes to the original code, with the intention of simplifying user experience and making it more widely deployable. These changes are:

1. Using sickle-trim (https://github.com/najoshi/sickle), a quality-based trimming package, due to its wide availability through all Python dependency managers (pip and Anaconda channels, including Bioconda);
2. Adapting the spacer extraction function to be parallelizable, which substantially reduces the computing time;
3. Reducing the number of intermediate files generated, and placing emphasis on data visualization;
4. Implementing an in-notebook calculation of the ordering scores.
5. Implementing a series of simulations meant to illustrate how scores would vary under different transcriptional programs, and giving users a starting point for interpreting their data and generating their own hypotheses.

2.3 Additional parameters to consider when designing experiments

2.3.1 Estimating initial sequencing depth

How accurately Retro-Cascorder's recording mirrors the true transcriptional activity of the bacterial population is mainly determined by the acquisition efficiency of Cas1-Cas2 and the number of cells used to reconstruct the recording. Each cell harbors one CRISPR array. Depending on the spacer sequence and promoter strength, around 0.5-5% of CRISPR arrays are expanded with a retron-derived spacer over the course of 24 hours. The majority of acquired spacers will instead either be from the bacterial genome or the plasmids harbored in the cell. These background spacers serve as a pseudo-internal timer and can even help to deduce the ordering of transcriptional activity of interest (see **Chapter 2.3.3**), assuming that background spacer integration is constant and independent from signal transcriptional activity. However, transcriptional order inference relies on informative CRISPR arrays, which we define as arrays that (a) contain at least two spacers, (b) contain at least one retron-derived spacer sequence, and (c) contain at least two spacer sequences are different from each other. Given both the rarity of retron-derived spacer acquisition and the exponentially diminishing probability that a given CRISPR array will contain more than one new spacer, many reads are often necessary to sequence enough informative arrays to calculate ordering scores. Furthermore, the greater the number of spacers, the higher the confidence will be that the scores reflect the true biological signal rather than estimation error from undersampling.

As a starting point, we suggest a sequencing depth of 1 million reads per biological sample, which is what we aimed for in our prior work for reconstructing the order of two discrete transcriptional events[21]. However, using a weaker promoter or shortening the duration of the recording would ultimately decrease the percentage of informative CRISPR arrays, which means a greater sequencing depth would be needed to find enough arrays to be used for ordering inference. **Fig. 2.3** shows how the accuracy of ordering scores calculated from simulated CRISPR arrays changes depending upon the number of informative arrays available.

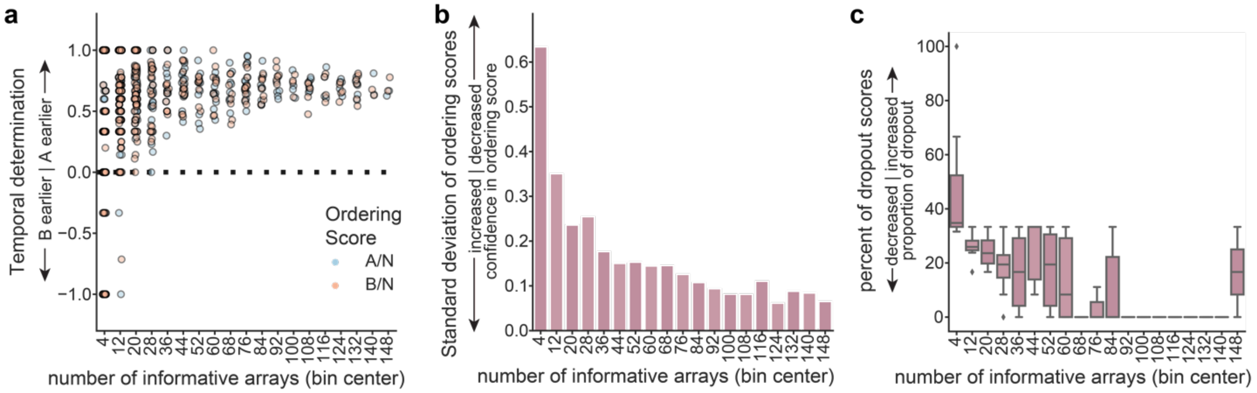


Figure 2.3. Plots summarizing the effect of the number of simulated informative arrays on ordering score accuracy. The X axis (number of informative arrays) is binned in groups of 8 and each label represents the center number of each bin. For clarity and to reflect both informative array sparsity and abundance, we simulated N=18 biological replicates for panels (a)-(c), with a range from 3,000 to 6 million arrays total. **(a)** Ordering scores calculated from simulated arrays which demonstrates how the number of informative arrays acquired over a simulated sequencing run impact the A/N and B/N ordering scores. As the number of informative arrays increase, the

exact magnitude and direction of the calculated ordering score converge towards its “true” value, suggesting more accurate ordering scores occur when there are higher numbers of informative arrays. **(b)** Effect of the increase in informative arrays on the standard deviation of A/N and B/N ordering scores. With increasing informative arrays, the standard deviation decreases, thus increasing the confidence in the ordering score. **(c)** Effect of the increase in informative arrays on the percentage of calculated A/N, B/N and A/B ordering scores that are “dropouts.” As the number of informative arrays increase, the percentage of “dropout” scores decreases, suggesting ordering information can be more reliably found when there are higher numbers of informative arrays. Error bars represent a 95% confidence interval.

The increased confidence in calling the order of transcriptional signals at deeper sequencing depths is primarily dependent on the number of informative arrays. Accordingly, we found that the standard deviation of the ordering scores decreased with higher numbers of informative arrays sequenced (**Fig. 2.3b**). Further, we found that, at low numbers of informative arrays, a majority of the ordering scores were “dropouts”, or incalculable, due to the paucity of observed A and B spacers used to calculate the scores (i.e., a “dropout” A/B ordering score would imply that there are no arrays with both A and B spacers sequenced) (**Fig 2.3c**). Based on these findings, we suggest that users will find that their recordings are most accurate when at least 40 informative arrays can be used for ordering calculations per biological replicate. In prior work[21] where we recorded the ordering of two strong promoters over 48 hours recording using a sequencing depth of 1 million reads, we regularly acquired over 40 informative arrays, and often closer to 100 informative arrays, per sample (10 of 12 biological replicates). However, since changes in the underlying transcriptional program, promoter strength, and/or leakiness may affect the sequencing depths necessary to observe enough informative arrays to enable robust calculation of ordering scores, we suggest that users empirically determine how many sequencing reads are necessary to reliably average around 40 informative arrays over several biological replicates. Generally, we consider increasing sequencing depth to be quickest and most affordable method to optimize the performance of Retro-Cascorder, before attempting to tweak other experimental variables, such as promoter strength.

Otherwise, if the user finds that the scores from their biological replicates are inconsistent or consistently result in dropouts after increasing sequencing depth, they can alternatively increase the number of biological replicates to account for the additional noise and variance. Another possibility is to use an approach like SENECA, described in another *Nature Protocols* paper[26], which enriches for expanded CRISPR arrays and thus would decrease the sequencing depth necessary to find informative arrays.

Although increased sequencing depth usually increases recording fidelity, some care has to be taken to ensure that sequencing depth is not *higher* than the number of original CRISPR sequences in the original sample. For this reason, we recommend that the number of genomes harvested be in large excess compared to the number of reads. Assuming perfect lysis and experimental conditions, we estimate that there should be no more than 400,000 unique genomes per sample, so we suggest users sequence using no more than 250,000 reads per sample, to be conservative. For example, to sequence at a depth of 1 million reads per biological replicate, we typically collect 4 separate samples from the same culture to prepare for sequencing, index them separately, then sequence each indexed sample at a depth of 250,000 reads before pooling all the reads together. By collecting multiple samples from the same culture, the ratio of the number of starting CRISPR arrays to the eventual number of sequencing reads is increased to minimize any risk of re-sequencing the same CRISPR array more than once.

2.3.2 Designing the signal plasmid

The signal plasmid, which can link up to two promoters of interest with barcoded retron noncoding RNA, is one of two essential plasmids for Retro-Cascorder. For users to clone a signal plasmid which contains their own promoters of interest, we recommend using our signal plasmid pSBK.134 as the backbone. This plasmid contains two inducible promoters which face in opposite directions to prevent expression leakage or crossover. Each promoter expresses one of two barcoded Eco1 noncoding RNAs.

The two barcodes on pSBK.134 were found to have similar acquisition efficiencies to each other. However, we have also previously tested other barcodes, which result in similar acquisition rates to those used in pSBK.134 (see Fig. 2b in our original publication[21] for the observed acquisition efficiency for each barcode) and can replace the barcodes used in pSBK.134 in case the user needs a different sequence. All barcode sequences whose acquisition efficiency have been previously validated are listed in **Table 2.1** below:

Table 2.1 Replacement barcodes for use in the signal plasmid

Number of barcode in Fig 2b of ref. 21	Barcode
1	CCTAGG
2	GCTAGC
3	CTGCAG
4	GTGCAC
5	ACGCGT
6	CAGTAG
7	GAGCTC
8	GCATGC

In cases where users would prefer a different plasmid backbone or architecture, they can also generate different plasmid designs in which a barcoded Eco1 noncoding RNA is under the control of a chosen promoter. However, we find that the choice of plasmid and backbone architecture alters acquisition rates. Other architectures may result in acquisition rates similar to the pSBK.134 architecture, but they must be tested first. Ideally, the expression from a retron noncoding RNA over 24 hours should result in CRISPR arrays expanded with new, retron-derived spacers at a rate of between 0.5-5%. If not, a different plasmid architecture should be picked or the promoter of interest may be too weak for Retro-Cascorder to accurately resolve temporal recordings that use it without increasing sequencing depth or the number of biological replicates.

2.3.3 Calculating and interpreting ordering scores

The exact formulas used in each of the three scores are shown below:

$$\mathrm{Score}_{\mathrm{A/N}} = \frac{\mathrm{f}_{\mathrm{A}\ \rightarrow\mathrm{N}\ \rightarrow\mathrm{Leader}} - \mathrm{f}_{\mathrm{N}\ \rightarrow\mathrm{A}\ \rightarrow\mathrm{Leader}}}{\mathrm{f}_{\mathrm{A}\ \rightarrow\mathrm{N}\ \rightarrow\mathrm{Leader}} + \mathrm{f}_{\mathrm{N}\ \rightarrow\mathrm{A}\ \rightarrow\mathrm{Leader}}}$$

$$\mathrm{Score}_{\mathrm{B/N}} = \frac{\mathrm{f}_{\mathrm{N}\rightarrow\mathrm{B}\rightarrow\mathrm{Leader}} - \mathrm{f}_{\mathrm{B}\rightarrow\mathrm{N}\rightarrow\mathrm{Leader}}}{\mathrm{f}_{\mathrm{N}\rightarrow\mathrm{B}\rightarrow\mathrm{Leader}} + \mathrm{f}_{\mathrm{B}\rightarrow\mathrm{N}\rightarrow\mathrm{Leader}}}$$

$$\mathrm{Score}_{\mathrm{A/B}} = \frac{\mathrm{f}_{\mathrm{B}\rightarrow\mathrm{A}\rightarrow\mathrm{Leader}} - \mathrm{f}_{\mathrm{A}\rightarrow\mathrm{B}\rightarrow\mathrm{Leader}}}{\mathrm{f}_{\mathrm{B}\rightarrow\mathrm{A}\rightarrow\mathrm{Leader}} + \mathrm{f}_{\mathrm{A}\rightarrow\mathrm{B}\rightarrow\mathrm{Leader}}}$$

Here, $\mathrm{f}_{\mathrm{A}\rightarrow\mathrm{B}\rightarrow\mathrm{Leader}}$ is the count of arrays that have the spacers ordered as $A \rightarrow B \rightarrow \mathrm{Leader}$.

Although the A/N score and B/N score can mathematically range from -1 to +1, under the assumption that N spacer acquisition is constant, the scores should not exceed |0.5|. More precisely, the average value of the A/N and B/N scores over *n* replicates should be |0.5|, with the actual scores for each replicate being normally distributed around it. The spread of the distribution around this average score, or how much the A/N and B/N scores deviates from |0.5|, is a reflection of biological noise and variability in the recording system, as well as any deviation from the assumption that "N" is a constant signal independent of "A" and "B". If this assumption does not hold, or the CRISPR arrays sequenced are sparse (i.e., few new retron-derived spacers observed and usable to calculate the ordering scores; see Box 1), the ordering scores are more likely to fall outside this range. Although individual replicates may fall beyond the expected values, the average of a large enough sample size should nonetheless fall within -0.5 to +0.5. Thus, if a user observes that the average score falls outside of this range, we suggest increasing the sequencing depth, and if the result remains, re-evaluating the assumption regarding uniform and continuous "N" spacer acquisition.

2.4 Materials

2.4.1 Reagents

2.4.1.1 General

- UltraPure^{TM} Distilled Water (Thermo Fisher Scientific, cat. no. 10977015)

2.4.1.2 Plasmid backbones

- Expression plasmid pSBK.079 (Addgene, cat. no. 187218)
- Signal plasmid pSBK.134 (Addgene, cat. no. 187219)

2.4.1.3 Biological Materials

- BL21-AI^{TM} One Shot^{TM} Chemically Competent *E. coli* (Thermo Fisher Scientific, cat. no. C607003) **CRITICAL** BL21-AI *E. coli* is widely available but contains an endogenous retron. We do not find that this endogenous retron interferes with Retro-Cascorder, but users can instead use a modified BL21-AI *E.coli* strain bSLS.114 that lacks this endogenous retron (Addgene, cat. no. 191530).

2.4.1.4 Recording experiment

- LB broth (Miller; 10g/L tryptone, 5g/L yeast extract, 10 g/L NaCl, UltraPure^{TM} Distilled Water, pH 7)

- L-arabinose (200 mg/mL in UltraPure™ Distilled Water, sterile-filtered; GoldBio, cat. no. A-300)
- IPTG (100 mM in UltraPure™ Distilled Water, sterile-filtered; GoldBio, cat. no. I2481C)
- Kanamycin (35 mg/mL in UltraPure™ Distilled Water, sterile-filtered; GoldBio, cat. no. K-120)
- Carbenicillin (100 mg/mL in 50% vol. UltraPure™ Distilled Water/50% vol absolute ethanol; GoldBio, cat. no. C-103) !CAUTION Carbenicillin is a respiratory and skin sensitizer; avoid breathing or skin contact. When handling carbenicillin, wear gloves and eye protection.
- Anhydrotetracycline (100 ng/mL in 50% vol. UltraPure™ Distilled Water/50% vol absolute ethanol, sterile-filtered; Cayman Chemical, cat. no. 10009542) !CAUTION Anhydrotetracycline is harmful if swallowed and causes skin and eye irritation. When handling anhydrotetracycline, wear gloves and eye protection.
- Choline chloride (100 µM in UltraPure™ Distilled Water, sterile-filtered; Sigma-Aldrich, cat. no. C7017)

2.4.1.5 Sample preparation for deep sequencing

- AMPure XP Reagent (Beckman Coulter, cat. no. A63881) **CRITICAL** To decrease cost, this reagent can also be substituted for Sera-Mag™ beads following a series of short washes and resuspension in homemade nucleic acid binding buffer. Refer to Supplementary Methods for details.

- Q5 High-Fidelity DNA Polymerase (NEB, cat. no. M0318L) **CRITICAL** Use of a high-fidelity polymerase minimizes errors in amplification of high-diversity libraries for multiplexed sequencing.
- Q5 Reaction Buffer 5x (NEB, cat. no. B9027S)
- Deoxynucleotides (dNTPs) Solution Mix, Nucleotide 10 mM, 40 µmol each nucleotide (New England Biolabs, cat. no. N0447L)
- SYBR Green I Nucleic Acid Gel Stain, 10,000X concentrate in DMSO (Thermo Fisher Scientific, cat. no. S7585)
- ROX Low Reference Dye (Kapa Biosystems, cat. no. KD 4601) **CRITICAL** This reference dye is used with the StepOnePlus Real-Time PCR machine mentioned in Equipment below. If using a different machine, use the appropriate reference dye as provided by the machine's instructions.
- 1 Kb Plus DNA Ladder (Thermo Fisher Scientific, cat. no. 10787018)

2.4.1.6 Deep sequencing

- KAPA Library Quantification Kit – Complete Kit (Universal) (KK4824, Roche cat. no. 07960140001)
- KAPA Library Quantification DNA Control Standard, Illumina (KK4906, Roche cat. no. 7960417001)
- PhiX Control Kit v3 (Illumina, cat. no. FC-110-3001)
- Miseq Reagent Kit v2 (300 cycle, Illumina, cat. no. MS-102-2002)
- Primers for deep sequencing (**Table 2.3**) **CRITICAL** All listed DNA primers can be purchased from DNA synthesis companies, such as IDT.

- 96-well PCR plate containing DNA indexing primer pair sets for MiSeq **CRITICAL** 96-well plates containing synthesized DNA primers can be purchased from DNA synthesis companies, such as IDT.

2.4.2 Equipment

- Plate Centrifuge (SouthwestScience, cat. no. SC20-PLATE)
- Adhesive PCR Plate Foils (Thermo Fisher Scientific, cat. no. AB0626)
- Bacterial culture tubes (VWR, cat. no. 60818-725)
- Water bath (VWR, cat. no. 1202)
- Bacterial shaker Innova S44i (Eppendorf, cat. no. S44I3100001)
- Disposable Pasteur Pipets, Flint Glass, 9” (VWR, cat. no. 14672-380)
- 1.5 and 2.0 mL Microcentrifuge tubes (Axygen, cat. no. MCT-150-C-S)
- Easy Reader Conical Polypropylene Centrifuge Tubes 15 and 50 ml (Thermo Fisher Scientific, cat. no. 07-200-886 and 05-539-8, respectively)
- 1-mm-gap cuvette (Bio-Rad, cat. no. 1652089)
- Benchtop microcentrifuge (Eppendorf, cat. no. 5425)
- Gene Pulser Xcell Electroporation System (Bio-Rad, cat. no. 1652666)
- PCR strip tubes, 0.2 ml (USA Scientific, cat. no. 102-4700)
- MiniAmp Plus Thermal Cycler (Thermo Fisher Scientific, cat. no. A37835)
- MicroAmp Fast Optical 96-Well Reaction Plate, 0.1 mL (Thermo Fisher Scientific, cat. no. 4346907)
- StepOnePlus Real-Time PCR System (Applied Biosystems, cat. no. 4376600)

- E-Gel EX agarose gels, 2% (Thermo Fisher Scientific, cat. no. G402002)
- E-Gel Power Snap Electrophoresis Device (Thermo Fisher Scientific, cat. no. G8100)

  CRITICAL In place of E-Gel EX agarose gels and E-Gel Power Snap Electrophoresis Device, a 2% agarose gel can be poured in the lab and run on a standard gel electrophoresis system. For more information on specific equipment and protocols to both make and run agarose gels, see https://www.jove.com/t/3923/agarose-gel-electrophoresis-for-the-separation-of-dna-fragments[37]
- Magnetic separator for 96-well PCR plate (DynaMag-96 Side Skirted Magnet; Thermo Fisher Scientific, cat. no. 12027)
- MiSeq system (Illumina, cat. no. SY-410-1003)

2.4.3 Software

- The computational methods described in this protocol have been implemented for a Unix-like operating system with a bash shell. This Jupyter-notebook (written in Python3) serves as a self-contained, interactive walkthrough of the deep sequencing data generated during our experiments, and requires *Jupyter-notebook* to be installed; the rest of the dependencies are handled internally. Note that the analysis pipeline is meant to be run on a Unix-like operating system; nonetheless, it can be adapted to run on Windows-based OSs with minimal changes to the notebook.

Python dependencies are listed below. For new python users, we strongly recommend beginning with the Anaconda Distribution (https://www.anaconda.com/distribution) - it includes Python and many other commonly used packages for scientific computing and data science.

Anaconda also enables easy installation of dependencies. See Troubleshooting for more information.

- *Python* ≥ v3.0, available here (https://www.python.org/downloads/)
- *Jupyterlab*, to run the notebook, available here (https://jupyter.org/install)[38]
- *Biopython*, a set of freely available tools for biological computation, available here (https://biopython.org/wiki/Download)
- *Fuzzysearch,* a python package for string matching, available here (https://pypi.org/project/fuzzysearch/)
- *sickle-trim*, a python package for read trimming, available here (https://github.com/najoshi/sickle)[39]
- *seaborn,* a data visualization library based on matplotlib, available here (https://seaborn.pydata.org/installing.html)
- *numpy,* a package for scientific computing, available here (https://numpy.org/install/)
- *pandas,* a powerful data analysis and manipulation tool, available here (https://pandas.pydata.org/getting_started.html)
- *matplotlib,* a comprehensive library for creating static, animated, and interactive visualizations, available here (https://matplotlib.org/stable/users/getting_started/)
- *multiprocess,* a library that enables multiprocessing and multithreading in python, available here (https://pypi.org/project/multiprocess/#files)

We recommend installing these packages through pip, or Anaconda's own package handler, with the following command prompts: "pip install jupyterlab biopython fuzzysearch seaborn numpy multiprocess pandas matplotlib; conda install -c bioconda sickle-trim".

2.4.4 Reagent Setup

2.4.4.1 Preparing primers for deep sequencing

Synthesize primers as single-stranded oligonucleotides; sequences are listed in **Table 2.3**. Dissolve each oligo in UltraPure^{TM} Distilled Water to a final concentration of 100 µM. To store, keep at -20°C for up to 1 year.

2.4.4.2 Stock 96-well plate of indexing primer (100 µM each)

Starting from a 96-well PCR plate containing a pair of dried forward and reverse indexing primers per well, spin plate in a plate spinner to collect dried primers at the bottom of each well. Add UltraPure^{TM} Distilled Water to each well to a final concentration of 100 µM per each primer in a well. To store, seal the 96-well PCR plate with an adhesive plate foil and keep at -20°C for up to 1 year.

2.5 Procedure

2.5.1 Temporal recording procedure; transformation of signal and expression plasmids into expression strain

● Timing 4 d, 1.5 h hands-on

1. Place one aliquot of *E. coli* BL21-AI cells on ice to thaw. When aliquot is fully thawed (5-10 min), transfer between 15-50 µL of cells per transformation into a clean, pre-chilled bacterial culture tube.

   CRITICAL STEP Commercially available BL21-AI *E. coli* contains its own endogenous retron, whose RT-DNA appears on a PAGE gel. However, we find that its presence will not prevent Retro-Cascorder from recording the transcriptional recordings described in this protocol. If the presence of the endogenous retron impacts the user's experiments, an alternate BL21-AI *E.coli* strain bSLS.114 that lacks this endogenous retron is available through Addgene (catalog #191530).

2. On ice, add 1 pg-100 ng signal plasmid (i.e. pSBK.134) DNA in 1-3 µL total volume to the tube containing BL21-AI and gently swirl solution with pipette tip. Keep on ice for 15 minutes.

   CRITICAL STEP The transfection order of plasmids does matter. If the expression plasmid is added first, Cas1-Cas2 may acquire spacers before the transcriptional recording experiment begins. To guard against this risk, we recommend always adding the expression plasmid last (see Step 10).

3. Heat shock cells by placing the tube in a water bath heated to 42°C for 30 s. Place the tube back on ice for 1 minute.
4. Add 250 µL SOC media to the tube and place in the bacterial shaker at 37°C at 250 r.p.m. for 1 hour to allow cells to recover and express antibiotic resistance gene.
5. Plate entire transformation on a pre-warmed LB agar plate (10 cm, 35 µg ml–1 kanamycin). Spread bacteria with flamed Pasteur pipet in a dilution series to promote formation of individual colonies (if unfamiliar with how to perform a dilution series, see https://www.jove.com/v/10507/serial-dilutions-and-plating-microbial-enumeration[40]).
6. Incubate LB agar plate overnight at 37°C for 16 hours.
7. The following morning, check the LB agar plate for the presence of bacterial colonies. Take LB agar plates from 37°C and leave them at 4°C or room temperature (20°C) until the evening.

PAUSE POINT LB agar plates containing colonies of BL21-AI containing signal plasmid can be kept 4°C for up to 1 week before transforming with the expression plasmid.

8. Add 3 mL of LB media containing kanamycin (35 μg/mL) into a bacterial culture tube. Inoculate one tube with one bacterial colony from the LB agar plate. Transfer tube to bacterial shaker set at 37°C at 250 r.p.m. overnight for 16 hours.
9. The following morning, dilute 60 μL of culture into a bacterial culture tube with 3 mL LB containing the antibiotic kanamycin. Place tube in bacterial shaker set at 37°C at 250 r.p.m. and let incubate for 2 h.
10. During the 2 h incubation, prepare a 10 pg/μL solution of expression plasmid by diluting the expression plasmid (i.e. pSBK.079) in water to a total volume of 50 μL in a microcentrifuge tube. Place the expression plasmid solution, a 1-mm-gap cuvette, a microcentrifuge tube, and a 15 mL conical tube filled with water on ice to pre-chill.
11. After 2 h, transfer 1 mL of culture from Step 9 to the pre-chilled microcentrifuge tube from Step 10. Centrifuge the microcentrifuge tube in a microcentrifuge at 4°C for 30 s at 10,000 x g to pellet the culture.

CRITICAL STEP After 2 h, culture should barely be cloudy. Electroporation efficiency will be lower for a denser culture.

CRITICAL STEP For Steps 11-14, the bacteria should be kept cold at 4°C and all steps performed quickly to increase acquisition efficiency and avoid as much cell death as possible.

12. Remove supernatant by pipetting and resuspend cells in 1 mL chilled water from Step 10. Centrifuge the microcentrifuge tube in a microcentrifuge at 4°C for 30 s at 10,000 x g to pellet the culture.
13. Repeat Step 12 two more times for a total of three washes. If the bacterial pellet becomes loose once cells are in water, the microcentrifugation duration can be increased to 1 minute.
14. Remove supernatant and resuspend cells in 50 µL of expression plasmid solution from Step 10. Transfer 50 µL of the mixture to a pre-chilled cuvette from Step 10.
15. Dry the cuvette using a paper towel and place in the electroporation system. Electroporate using the following parameters: 1.8 kV, 25 µF, and 200 **Ω**.

    CRITICAL STEP The time constant (τ), indicating the time it takes for the voltage to decay to 1/3 the initial set voltage in milliseconds, following electroporation ideally should be between 4.7-5.2 and can be found on the display screen of the electroporation system after each electroporation. If any time constant is below 4, we recommend discarding the electroporation and trying again.
16. Quickly recover the electroporated cells into LB by pipetting 250 µL of SOC media into the cuvette and mixing with the cells. After, transfer the mixed solution from the cuvette into a bacterial culture tube. Place the tube in the bacterial shaker at 37°C at 250 r.p.m. for 1 hour to allow cells to recover and express antibiotic resistance gene.
17. Plate entire transformation on a pre-warmed LB agar plate (10 cm, 100 µg ml–1 carbenicillin and 35 µg ml–1 kanamycin). Spread bacteria with flamed Pasteur pipet in a dilution series to promote formation of individual colonies.

18. Repeat Steps 6-7. There should be bacterial colonies containing both the signal and expression plasmid on the LB agar plate.

    PAUSE POINT LB agar plates containing colonies of BL21-AI containing expression and signal plasmid can be kept at room temperature for up to 24 hours or at 4°C for up to 3 weeks before starting the recording experiment.

2.5.2 Temporal recording procedure; recording transcriptional activity for 2 days

● Timing 2 d, 40 min hands-on

19. For each biological replicate, add 3 mL of LB media containing carbenicillin (100 µg/mL) and kanamycin into a bacterial culture tube. Inoculate each tube with one bacterial colony from the LB agar plate from Step 18. Transfer tubes to bacterial shaker set at 37°C at 250 r.p.m. overnight for 16 hours.
20. The following morning, dilute 150 µL of culture into a bacterial culture tube with 3 mL LB containing the antibiotics carbenicillin and kanamycin and the inducers IPTG (1 mM) and l-arabinose (2 mg/mL) to induce expression of Cas1-Cas2. If appropriate, add a relevant compound (i.e. 3 µL of 100 ng/mL anhydrotetracycline to turn on pTet* promoter in pSBK.134) to induce expression of the first transcriptional event on the signal plasmid. Place tube in bacterial shaker set at 37°C at 250 r.p.m. and incubate for 8 h.
21. After 8 h, dilute 60 µL of culture into a new bacterial culture tube with 3 mL LB containing the same antibiotics and inducers as step 20. Transfer tubes to bacterial shaker set at 37°C at 250 r.p.m. and incubate overnight for 16 hours. CRITICAL STEP Bacteria are diluted after 8 h of growth to prevent bacteria from reaching stationary phase and to

allow cells to continue log-based growth. Additionally, always add new, fresh inducers to ensure continual expression of Cas1 and Cas2.

22. Repeat Steps 20 and 21, except—if appropriate—add a relevant compound (i.e. 30 µL of 100 µM choline chloride to turn on pBetI promoter in pSBK.134) to induce expression of the second transcriptional event on the signal plasmid.
23. After 48 h of bacterial growth and recording, collect 25 µL sample and mix with 25 µL water in a PCR tube. Boil at 95°C in a thermocycler for 5 min to lyse cells then allow to cool on benchtop (~5 min) before freezing at -20°C for later analysis.

PAUSE POINT Boiled bacterial samples can be stored at -20°C for at least 6 months.

2.5.3 Preparing CRISPR arrays for sequencing: Determine appropriate cycle number for first round PCR amplification

● Timing 2 h, 2h hands-on

CRITICAL: During first round PCR amplification, CRISPR arrays are amplified from the bacterial genome. Ideally, the PCR should stop during the end of log-based amplification before reaching the plateau. Minimizing excess numbers of cycles is important to reduce the potential for crossover events during the later cycles of a PCR. For each experimental paradigm or different signal plasmid, we recommend performing at least one qPCR amplification to determine the ideal cycling number to stop the PCR. After performing this step once, it should not have to be repeated for a given signal plasmid unless the user is attempting to troubleshoot potential issues downstream.

24. Thaw frozen bacterial samples from Step 23.

25. Dilute SYBR Green I in water to a final concentration of 5X for qPCR amplification. Given the size of the dilution, we recommend performing a serial dilution, i.e. add 1 µL SYBR Green I to 99 µL water and mix well. Then, add 15 µL of the mixture to 285 µL of water and mix well.

26. Prepare first round PCR primer mix by combining primers (see **Table 2.2**) as follows:

Table 2.2 Components for first round PCR primer mix

Component	**Amount** (µL)	**Final concentration** (µM)
H2O	90	-
SPCR_MiSeq3_fow1, 100 µM	2	2
SPCR_MiSeq3_fow2, 100 µM	2	2
SPCR_MiSeq3_fow3, 100 µM	2	2
SPCR_MiSeq3_fow4, 100 µM	2	2
SPCR_MiSeq3_fow5, 100 µM	2	2
Total	100	-

PAUSE POINT: Aliquots of first round PCR primer mix can be stored at -20°C for at least one year.

Table 2.3 Primer sequences required for deep sequencing

Primer name	Nucleotide sequence (5' to 3')	Purpose	Procedure step
SPCR_MiSeq3_fow1	CTTTCCCTACACGACGCTCTTCCGATCTNCATTAATTAATAATAGGTTATGTTTAGAGTGTTCC	First round PCR	26, 31
SPCR_MiSeq3_fow2	CTTTCCCTACACGACGCTCTTCCGATCTNNCATTAATTAATAATAGGTTATGTTTAGAGTGTTCC	First round PCR	26, 31
SPCR_MiSeq3_fow3	CTTTCCCTACACGACGCTCTTCCGATCTNNNCATTAATTAATAATAGGTTATGTTTAGAGTGTTCC	First round PCR	26, 31
SPCR_MiSeq3_fow4	CTTTCCCTACACGACGCTCTTCCGATCTNNNNCATTAATTAATAATAGGTTATGTTTAGAGTGTTCC	First round PCR	26, 31
SPCR_MiSeq3_fow5	TTTCCCTACACGACGCTCTTCCGATCTNNNNNCATTAATTAATAATAGGTTATGTTTAGAGTGTTCC	First round PCR	26, 31
SPCR_MiSeq3_rev	GGAGTTCAGACGTGTGCTCTTCCGATCTGTGTCAACAATCGTTCCCTGATTGTC	First round PCR	26, 31
P5	AATGATACGGCGACCACCGA	Indexing PCR	37
P7	CAAGCAGAAGACGGCATACGAGAT	Indexing PCR	37

CRITICAL STEP Since the Illumina MiSeq system calibrates on a diverse set of nucleotides, we recommend using a mixture of at least 5 forward primers with varied lengths and nucleotides to ensure accuracy of the subsequent reads, assuming that the majority of the run will include CRISPR arrays. In the case that a diverse set of amplicons will be run alongside the arrays, there is no need to use varied forward primers.

27. Prepare the qPCR reaction according to **Table 2.4** on ice. We recommend creating a master mix with all reagents except DNA template. Dispense 24.2 µL of master mix per well then add 0.8 µL DNA template per well.

Table 2.4 Components for qPCR

Component	**Amount per reaction (µL)**	**Final concentration**
Q5 Reaction Buffer (5X)	5	1X
dNTPs (10 mM)	0.5	200 µM
Forward Primer mix from Step 26 (10 µM)	1.25	0.5 µM
SPCR_MiSeq3_rev (Table 1) (10 µM)	1.25	0.5 µM
Template DNA from Step 24	0.8	-
Q5 High-Fidelity DNA Polymerase	0.25	-
SYBR Green I from Step 25 (5X)	5	1X
H2O	10.95	-
Total	**25**	**-**

28. Begin qPCR reaction by implementing the following qPCR protocol according to **Table 2.5**:

Table 2.5 Protocol for qPCR

Cycle number	Denature	Anneal	Extend
1	98°C, 3 min		
2-46	98°C, 20 s	60°C, 15 s	72°C, 20 s

29. Note when during cycles 2-46 sample traces begin to plateau. This cycle number should be the number of cycles used during the subsequent first round PCR amplifications for all biological replicates with the same signal plasmid and experimental paradigm.

2.5.4 Preparing CRISPR arrays for sequencing: Amplifying and indexing samples

● Timing 6 h, 4 h hands-on

CRITICAL: This part of the Procedure amplifies the BL21-AI CRISPR array within the genome and attaches a primer extension that will be used for indexing.

30. Thaw frozen bacterial samples from Step 23.

31. *First round PCR amplification.* Prepare the PCR reaction according to **Table 2.6** on ice. We recommend first preparing a master mix without the template DNA. Dispense 24.2 µL of master mix per well then add 0.8 µL boiled bacterial sample per well.

Table 2.6 Components for first round PCR amplification

Component	**Amount per reaction (µL)**	**Final concentration**
Q5 Reaction Buffer (5X)	5	1X
dNTPs (10 mM)	0.5	200 µM
Forward Primer mix from Step 26 (10 µM)	1.25	0.5 µM
SPCR_MiSeq3_rev (Table 1) (10 µM)	1.25	0.5 µM
Template DNA from Step 30	0.8	-
Q5 High-Fidelity DNA Polymerase	0.25	-
H2O	15.95	-
Total	**25**	**-**

32. Perform PCR reaction using **Table 2.7**:

Table 2.7 Protocol for first round PCR amplification

Cycle number	Denature	Anneal	Extend
1	98°C, 30 s		
2-[cycle number determined in Step 29]	98°C, 10 s	72°C, 30 s	72°C, 30 s
[cycle number determined in Step 29] + 1			72°C, 2 min

Following PCR amplification, freeze each PCR reaction at -20°C or continue immediately to indexing.

CRITICAL STEP We typically find that there is no need to purify or clean-up the resulting products from the first round of PCR before moving forward to second round

PCR amplification. However, if the user finds that they do not acquire the expected product during the indexing reaction, a DNA clean-up may help during troubleshooting. PAUSE POINT First round PCR reactions can be stored at -20°C for at least 6 months.

33. Run samples out on a 2% agarose E-gel EX by loading 3.5 µL first round PCR product and 16.5 µL water into each well. Use the 1 kB+ ladder as a reference by adding 2 µL undiluted ladder and 18 µL water to the marker lane. Run gel for 10 minutes. Validate that the brightest PCR band is 265 nt. This band corresponds to the unexpanded CRISPR array, although higher-level bands may also be visible. These bands should be some multiple of 61 nucleotides larger than the unexpanded array, and correspond to expanded CRISPR arrays containing 1 or more additional spacers.

    CRITICAL STEP As mentioned in the Materials section, an alternative option to the 2% agarose E-gel EX is to create and run an agarose gel made in lab, as described in https://www.jove.com/t/3923/agarose-gel-electrophoresis-for-the-separation-of-dna-fragments[37]. For a homemade 2% gel, we recommend running the gel for around 30 min at 100V.

34. Create a 10 µM stock indexing plate from the 100 µM stock indexing plate (see Reagent Setup). We recommend first spinning the 100 µM stock indexing plate in a plate spinner to collect liquid at the bottom of each well before mixing 20 µL of each original 100 µM primer solution with 180 µL water in a new 96-well PCR plate. Afterwards, make a working plate at 100 nM by adding 2 µL of primer solution from the 10 µM stock plate with 198 µL water in another 96-well PCR plate. Both stock and working plates can be stored at -20°C for at least 1 year.

CRITICAL STEP Be careful not to cross-contaminate primers between wells by always spinning the plate to collect liquid before opening and never reusing pipette tips. In case of unexpected results downstream, making a new working plate is always safest.

35. Dilute DNA template for indexing by adding 5 µL first round PCR reaction from Step 32 into water for a total volume of 75 µL. Pipette up and down vigorously to mix the reaction.
36. Dilute SYBR Green I in water to a final concentration of 5X for qPCR amplification. Given the size of the dilution, we recommend performing a serial dilution, i.e. add 1 µL SYBR Green I to 99 µL water and mix well. Then, add 15 µL of the mixture to 285 µL of water and mix well.
37. *Indexing.* This step indexes each amplicon with a unique, sample-specific barcode necessary for subsequent deep sequencing. Prepare the qPCR reaction according to **Table 2.8** on ice. We recommend creating a master mix with all reagents except DNA template and the forward and reverse P5 & P7 indexing primers (**Table 2.3**). Dispense 23 µL of master mix per well then add 1 µL of each indexing primer and 5 µL DNA template per well.

Table 2.8 Components for indexing qPCR

Component	Amount per reaction (µL)	Final concentration
H2O	12.18	-
Q5 Reaction Buffer (5X)	6	1X
SYBR Green I from Step 36 (5X)	3	0.5X
dNTPs (10 mM)	0.9	300 µM
P5 primer (25uM)	0.36	0.3 uM
P7 primer (25uM)	0.36	0.3 uM
Q5 Hotstart Polymerase	0.6	-
Rox (50X)	0.6	1X
Indexing primers (F & R) from Step 34 (100 nM each)	1	3.33 nM each
DNA template from Step 32	5	-
Total	30	-

CRITICAL STEP Do not substitute a high-fidelity DNA polymerase like Q5 with a qPCR polymerase. Use of a qPCR polymerases may result in higher error rates which decrease the fidelity and accuracy of reads in the deep sequencing library.

CRITICAL STEP This reaction is modified from a normal PCR to include two sets of distinct primers rather than a single primer set. The indexing primers are very long and have a propensity to create unwanted products such as primer dimers, so they are added at a low concentration. Meanwhile, the P5 and P7 primers are added at a higher concentration. The intention is to add indices based on the indexing primers to the amplicons in the initial cycles and then to avoid later unwanted products by allowing the P5 and P7 primers to amplify the indexed molecules throughout the rest of the cycles.

38. Begin qPCR reaction by implementing the qPCR protocol in **Table 2.9**:

Table 2.9 Protocol for indexing qPCR

Cycle number	Denature	Anneal	Extend
1	98°C, 3 min		
2-46	98°C, 20 s	60°C, 15 s	72°C, 20 s

39. If sample amplification traces begin to approach plateau during cycles 2-46, stop machine and move 25 µL of these samples to a new, clean 96-well PCR plate. Preferably, samples should be moved before their traces reach plateau phase, to avoid overamplification and subsequent artifacts. We recommend removing the samples as the amplification curves begin to flatten, which ideally corresponds to 2-3 cycles before the plateau.

    Make a new qPCR run with the same protocol as Step 38 and restart the original plate with the remaining samples.

40. Repeat Step 39 until all indexed samples have plateaued and been collected into the same 96-well plate. Continue immediately to DNA clean-up or seal the 96-well PCR plate with an adhesive plate foil before freezing at -20°C.

    PAUSE POINT Indexed amplicons can be stored at -20°C for at least one month.

    CRITICAL STEP Collecting indexing product before plateauing minimizes the chance for chimeric amplicons and index swapping. However, to save time, we do not recommend stopping, removing samples, and restarting a plate more than two times. To avoid too many restarts, we recommend pulling the whole plate after multiple samples have plateaued and choosing to collect samples in batches, rather than waiting for and collecting each individual sample as it leaves its exponential phase.

41. *PCR clean-up using beads.* This step purifies indexed amplicons without biasing the amplicons based on their size using beads. To save on reagent costs, we typically prepare and wash our own SPRI beads. Refer to **Chapter 2.9** for details. However, to save time,

commercially available XP AMPure beads may also be used interchangeably without any additional preparations or wash steps. If using XP AMPure beads, add a volume of beads at a 1.8X bead-to-DNA dilution to each well in the 96-well plate containing indexed product. For example, for a 1.8X bead-to-DNA dilution, add 45 µL beads to 25 µL indexed product per well. Otherwise, if using homemade SPRI beads, determining the optimal ratio of beads to add is explained in **Chapter 2.9**.

42. Mix reaction thoroughly by pipetting between 10-15 times. Incubate reaction for 5 minutes at room temperature. Place 96-well plate on magnet until beads migrate near the magnet and the solution is clear.
43. Remove the supernatant by pipetting and discard.
44. Wash the DNA by adding 200 µL fresh 70% ethanol and allow to incubate on the magnet for one minute until the solution is clear. Remove supernatant by pipetting and discard.
45. Repeat Step 44 to wash DNA one more time.
46. Let DNA dry for <3 minutes.

    CRITICAL STEP DNA needs to dry to remove contaminant ethanol but allowing the beads to dry for too long will result in low recovery. The presence of cracks appearing in the beads is a sign that the DNA has been allowed to dry for too long.

47. Remove the plate from the magnet and resuspend the DNA with 25 µL water. Pipette up and down between 10-15 times to mix. Incubate at 5 minutes at room temperature.
48. Place the plate back on the magnet and wait until the solution is clear. Collect supernatant and move into a new, clean 96-well PCR plate. Continue immediately to deep sequencing or seal the 96-well PCR plate with an adhesive plate foil before freezing at -20°C.

PAUSE POINT Purified, indexed samples can be stored at -20°C for at least a month.

49. Run samples out on a 2% agarose E-gel EX by loading 3.5 μL indexed product and 16.5 μL water into each well. Use the 1 kB+ ladder as a reference by adding 2 μL undiluted ladder and 18 μL water to the marker lane. Run gel for 10 minutes. Validate that the brightest PCR band is 402 nt. This band corresponds to the unexpanded CRISPR array, although higher-level bands may also be visible. These bands should be some multiple of 61 nucleotides larger than the unexpanded array, and correspond to expanded CRISPR arrays containing 1 or more additional spacers.

2.5.5 Multiplexed sequencing of CRISPR arrays

● Timing 3.5 h, 2 h hands-on

50. Dilute cleaned-up and indexed samples from Step 48 1:40,000 in water for quantification. We recommend performing this step through serial dilution, i.e. add 1 μL sample from Step 48 to 499 μL of water and mix well. Then, add 10 μL of the mixture to 790 μL of water and mix well.

51. Set up qPCR to quantify amount of each diluted sample using the KAPA Library Quantification Kit along with their DNA Control Standard. Each standard should be run in duplicate. In a 96-well plate, prepare reactions using **Table 2.10** on ice:

Table 2.10 Components for quantification qPCR

Component	Amount per reaction (µL)
KAPA Mastermix (with primers and ROX added previously, according to manufacturer's instructions)	6.2
1:40,000 diluted sample from Step 50 OR undiluted standard	4

52. Run qPCR according to the protocol included with the KAPA Library Quantification kit.
53. Using qPCR results, calculate molar concentrations of cleaned-up, indexed samples using the KAPA Library Quantification Data Analysis Template provided by the KAPA Library Quantification Kit.
54. After determining the molarity of each indexed sample, normalize the samples by adding the appropriate volume of each sample, along with water, to a single microcentrifuge tube to produce a multiplexed library that yields the desired number of reads for each sample. We recommend using the software tool "Pipette-Guide-96" (https://github.com/tamilieberman/Pipette-Guide-96) when pipetting samples to help save time and keep track of the work.
55. Dilute and denature multiplexed library according to the MiSeq System Denature and Dilute Libraries Guide (https://support.illumina.com/sequencing/sequencing_instruments/miseq/documentation.html) from Illumina.

    CRITICAL STEP Due to the low diversity of templates present in the library, use a PhiX control spike-in of 10% as directed on page 10 of the MiSeq System Denature and Dilute Libraries Guide. Libraries with low diversity often have unbalanced nucleotide composition, or the relative proportion of each of the four nucleotide bases. In such cases, the MiSeq instrument may fail to accurately sequence the samples. To compensate, the PhiX control spike-in provides a more balanced base composition to improve the sequencing run quality.

56. Prepare the sample sheet for the MiSeq run with the appropriate information, such as chemistry, number of cycles, and indices.
57. Load the MiSeq instrument and run. For additional information on deep sequencing, we recommend referencing the MiSeq System Guide from Illumina.

2.5.6 Data analysis

● Timing 1-3 h (depending on number of CPU cores available), 30 min hands-on

CRITICAL We have adapted the scripts pertaining to our original publication, available on our GitHub (https://github.com/Shipman-Lab/Spacer-Seq), into a Jupyter notebook[38] (https://docs.jupyter.org/en/latest/), written in Python. This notebook serves as a self-contained, interactive walkthrough of the deep sequencing data generated during our experiments and can also be used by users analyzing their own code by running each notebook cell in order when using their FASTQ files from the deep sequencing run in Step 57. The notebook requires JupyterLab or a similar ipython-notebook handler to be installed; the rest of the dependencies are handled internally within the notebook. Note that the analysis pipeline is meant to be run on a Unix-like operating system; nonetheless, it can be adapted to run on Windows-based OSs with minimal changes to the notebook, which are pointed out in the notebook where relevant. This notebook focuses on recreating figure 4L from our original publication[21], which shows the ordering analysis of recording experiments with signal plasmid pSBK.134 (as detailed in Steps 58-73). Hence, the data downloaded will be that pertaining to figure 4L from ref. 21.

58. If necessary, using terminal, install JupyterLab with pip.

59. Download the GitHub repository (https://github.com/Shipman-Lab/Spacer-Seq_Nat-Protocols). The simplest way is to download it as a .zip file and uncompress it.

60. Using terminal, go to the GitHub repository directory: "cd Spacer-Seq_Nat-Protocols-main".

61. The notebook uses the following dependencies:

    a. fuzzysearch

    b. Biopython

    c. seaborn

    d. numpy

    e. sickle-trim

    These can be installed by running a cell in the notebook which verifies that the required dependencies are installed, or installs them if need be.

62. Import the necessary Python packages and dependencies.

63. Run the relevant "Step 63" cell in the notebook to load a dataframe with metadata relevant to the FASTQ files that the user wants to analyze.

    (A) If users would like to use example FASTQ files we've provided to recreate Fig. 4L from ref. 21, our Jupyter notebook is set up to load the relevant Sequence Read Archives (SRA) run table by running the "Step 63(a)" cell, which contains metadata describing the sequencing files. This file is provided in the GitHub repository with the notebook, but can be accessed through the NCBI Sequence Read Archive (PRJNA838025).

(B) Alternatively, if users would like to analyze their own FASTQ files from the sequencing run in Step 57, they can perform the following steps:

(i) Download the FASTQ files from the sequencing run in Step 57, and save the FASTQ files into a directory called “fastqs” located in the same directory as the Jupyter notebook.

(ii) Users should create a spreadsheet containing necessary metadata about their samples. This file should be a tab-delimited, spreadsheet-style table with columns that include "Library Name", "Condition", "Replicate", "PCR", and "Order". Column "Library Name" should contain the name of the FASTQ file to be analyzed (e.g., "msSBK-2-35_S35_L001_R1_001.fastq.gz"); "Condition" is a description of the experiment run (e.g., "BA_PCR2"); "Replicate" is the biological replicate (e.g., 1); "PCR" is the technical PCR replicate (e.g., 3); and "Order" is the order of the experiment run (e.g., "AB" for an experiment where signal "A" is expected to have been present before signal "B"). We recommend creating the file in Microsoft Excel with the aforementioned columns and saving the output as a .txt file.

(iii) Save the metadata spreadsheet as “SraRunTable.txt” in the same directory as the Jupyter notebook.

(iv) Run the “Step 63(b)” cell in the notebook to load the metadata dataframe.

CRITICAL STEP There are a number of ways that the user can retrieve FASTQs from previous sequencing runs available through the NCBI SRA. This step can be performed manually. However, we recommend using `SRA-tools`, a collection of tools and libraries, developed by

NCBI for the purpose of interacting with the SRA. This collection allows reasonably quick querying and downloading of the FASTQs. Of note, the most recent release of `SRA-tools` is not available through `pip` or Python's usual dependency managers. Instead, it should be installed manually and interactively. To circumvent this, we have written a snippet of code that allows users to download the most recent release of `SRA-tools` and use its packages locally. With this, users can specifically query and download the FASTQ files relevant to the analysis to be performed (i.e., the data pertaining to figure 4L of ref. 21).

CRITICAL STEP The snippet of bash code used to download and run `SRA-tools` is written for a Unix-like OS – in the notebook, we have illustrated how to adapt it to run on MacOSX, and have suggested how users can adapt this to work on other OSs.

64. Run the "Step 64" cell in the notebook to trim the FASTQs using `sickle-trim`, a Python package that uses sliding windows along with quality and length thresholds to determine when quality is sufficiently low to trim the 3'-end of reads and also determines when the quality is sufficiently high enough to trim the 5'-end of reads[39].

65. Run the "Step 65" cell in the notebook to set global variables, such as the "Repeat" sequence and the "old" spacers' sequences. These are the spacers found in the CRISPR array of BL21AI *E. coli*. Additionally, we define how stringently the query sequences have to match the references (i.e., how closely a putative repeat has to match the actual repeat sequence). This allows some tolerance for sequencing errors. By default, we set the repeat fuzziness to 4 (i.e., allowing 4 mismatches between query and reference) and the old spacers fuzziness to 5.

66. Run the “Step 66” cell to define the following functions that will be used for the analysis. These functions perform most of the analysis, and work as follows:

    `get_spcrs(sequence)`: takes as input a sequence (typically a single read), and returns a list of spacers extracted from said read.

    `not_existing(spacer)`: takes as input a sequence (typically a putative spacer), and determines whether this sequence resembles (≥83% similar) an old spacer or a repeat. Returns `False` if so; if not, returns `True` -- this is how new spacers (i.e, the results of new CRISPR array expansions) are identified.

    `get_spcrs_11BC(sequence)`: takes as input a sequence (typically a single read), and returns a list of spacers extracted from said read. This function works analogously to `get_spcrs`, with one important difference, as described in our original publication[21]. `*get_spcrs_11BC* is an implementation of the “lenient analysis” used in Figs. 4 and 5 of ref. 21, where a retron-derived spacer was defined to be a spacer that contained an 11-base region of the hypothetical prespacer consisting of the 7-base barcode region and 2 bases on either side (with one mismatch or indel allowed).

    `matchesTarget(target, seq)`: takes as input a target and reference sequence, and returns `True` if the sequences are the same (with an allowance of 1 mismatch or change); returns `False` otherwise.

    `double_order(double)`: takes as input two spacers from a double expansion, and returns a tuple of coded spacers, e.g. (‘A’, ‘B’) or (‘B’, ‘N’).

    `triple_order(triplet)`: takes as input three spacers from a triple expansion, and returns a tuple of coded spacers, e.g. (‘A’, ‘B’, ‘N’).

` *multiprocess_spr(file)* `: this function will:

- setup a temporary dictionary, `ddd`, to store the new spacer data;
- generate a counter of the reads in the input FASTQ, for the sake of expediting the analysis;
- iterate through each read in the counter, extract and and determine the characteristics of the read and its spacer(s), such as:
    - o does the read contain one or more spacers;
    - o are the spacers "old" (one of the spacers found in the endogenous CRISPR array) or "new";
- store the read and spacer information in the temp dictionary `ddd` as a dictionary ~`{"FASTQ_i": ddd}`, where `ddd` is the dictionary with the information collected on all of the FASTQ reads;
- return the dictionary for downstream analysis.

Note that the function called to extract the spacers is *`get_spcrs`*, which takes as input a read, and outputs a list of spacers. This list of spacers is then processed by the rest of the *`multiprocess_spr`* function and the features detailed above are extracted and used to bin the spacers and reads, which are finally added to the temporary dictionary `ddd`, as discussed above.

67. For each read in each FASTQ, run the "Step 67" cell to extract new spacers and store them according to their characteristics and the characteristics of the CRISPR arrays from which they were extracted. The idea is to execute the function defined above as *`multiprocess_spr`*, which relies on two functions defined in Step 66 as follows:

- Uses '*get_spcrs'* to extract spacers from each read

- Uses ‘*not_existing*’ to check whether an extracted spacer is an old, prexisting spacer in the array (to qualify, the spacer has to be ≥83% similar to an old spacer) or a repeat. If a spacer meets neither criterion, it is instead considered a new spacer that was acquired over the course of the experiment.

To speed things up, this analysis uses multiprocessing to offload tasks to worker processes, and enables the analysis of multiple FASTQs in parallel. The number of processes run will be `cpu_count - 1`, where `cpu_count` is the number of CPUs in the system (i.e., on your laptop or cluster).

68. Store the data collected (information about of FASTQs, their reads, and spacers) in a dictionary by running the “Step 68” cell. If users are using the example FASTQ files provided in Step 63(a), this dictionary, `dict_data`, will contain a lot of useful information, most of which will not be used to re-create Fig. 4L from ref. 21, but can be explored by users.
69. Determine the order of spacers in each sequenced CRISPR array by running the “Step 69” cell. This cell works by running the following functions:
    - ‘*get_spcrs_11BC’* to extract potential retron-derived sequences from spacers in each, similarly to the *‘get_spcrs’* function. However, this function defines a retron-derived spacer as one that contains an 11-base region of the hypothetical prespacer, consisting of the 7-base barcode region, and 2 bases on either side (with one mismatch or indel allowed). For instance, an “A” retron-derived spacer would have an 11bp core region consisting of the following sequence: “GTTGCAGCAAC”. Similarly, a “B” retron-derived spacer would have an 11bp core region consisting of the following sequence: “GTCAGACTGAC”.

- ‘*matchesTarget’* to determine if the potential retron-derived spacer sequences are “A”, “B”, or “N” spacers, as specified in the ‘Target_dict.’
- ‘*double_order’* and ‘*triple_order’*, which iterates through every FASTQ, generating a dictionary of the counts of every possible permutation of “ABN” spacers, both for double expansions and triple expansions. For instance, in the case of double expansions, the possibilities are:
    - A, A
    - A, B
    - A, N
    - B, B
    - B, A
    - B, N
    - N, N
    - N, A
    - N, B

These counts are stored in the dictionaries `double_dict` and `triple_dict`. Note that the function called is `get_spcrs_11BC`, because it involves a more 'relaxed' search for retron-derived spacers, as mentioned above.

70. Run the “Step 70” cell to generate a dataframe with the data collected in Step 69. Specifically, the code in this cell generates a dataframe `ordering_df' by merging the dictionaries of double and triple spacer expansion ordering counts created in Step 69. Then, the code merges the ‘ordering_df’ dataframe with the metadata dataframe generated in Step 63. This cell also adds two columns to this new dataframe called `Order`, or what the

experimental order of signals were (A → B or B → A), and 'PCR', which will allow us to average scores within biological replicates.

71. Run the "Step 71" cell to sum the number of informative arrays (i.e., (A, N), (A, B) ...) for each biological replicate, which is stored in the `summed_counts` dataframe.

72. Calculate the "Ordering Scores" by running the "Step 72" cell. The A/N score is calculated by subtracting the total number of (A, N) arrays from the total number of (N, A) arrays, then dividing that value by the sum of the total number (A, N) and (N, A) arrays. The B/N score is calculated by subtracting the total number of (N, B) arrays from the total number of (B, N) arrays, then dividing that value by the sum of the total number of (N, B) and (B, N) arrays. The A/B score is calculated by subtracting the total number of (A, B) arrays from the total number of (B, A) arrays, then dividing that value by the sum of the total number of (A, B) and (B, A) arrays. As discussed in the Experimental Design section and Box 3, these logical rules should govern the ordering of spacers in the CRISPR arrays and assist with inferring the order of transcription of tagged genes (in this case, of distinct ncRNAs).

CRITICAL STEP Because spacers are acquired unidirectionally, with newer spacers closer to the leader sequence, we propose that, if transcript "A" is expressed before transcript "B", A → B → Leader arrays should be more numerous than B → A → Leader arrays. Conversely, if "B" is expressed before "A", the number of B → A → Leader arrays should be greater than the number of A → B → Leader arrays. For a more extensive discussion of the scores, refer to "Analysis" section.

73. To visualize the data, the ordering scores for a given experiment can be plotted as a strip or swarm plot. We also recommend users add a horizontal line at 0 to separate scores

corresponding to "A happened before B" (positive ordering score values) from scores corresponding to "B happened before A" (negative ordering score values). An example of such a plot is provided in Anticipated Results (see Fig. 4) and how to generate such a plot from the calculated ordering scores is shown in the two cells that follow the “Step 73” heading. Users first generate a smaller dataframe, `summarized_df`, that contains information about the filename, the order, the biological replicate, the PCR, the score, and type of score (i.e. A/N, B/N, or A/B). Afterwards, the ‘seaborn’ package is used to generate two overlaid plots:

- A swarmplot, showing the mean value of each score per biological replicate;
- A violinplot, to give a sense of the distribution of the scores.

CRITICAL: This is the end of the pipeline to calculate ordering scores for two transcriptional events and create plots similar to Fig. 4L of ref. 21 (see also **Fig. 2.4**).

2.6 Troubleshooting

Troubleshooting advice can be found in **Table 2.11**.

Table 2.11 Troubleshooting steps

Step	Problem	Possible reason	Solution
15	Low time constants following electroporation; arcing during electroporation	Solution is too conductive, potentially due to too much salt or DNA.	Ensure solution is salt-free by performing extra washes and removing all supernatant during wash steps; decrease the concentration of DNA added to the cuvette
22, 23	Bacteria not growing; low culture density or OD	Inducible compounds or signal plasmid inhibits growth	Optimize dilution amount while passaging or length of transcriptional recording to allow bacteria to near stationary phase before passaging or collecting for harvest
29, 39	Trace shows no clear log-based amplification; slope is very shallow throughout the entire qPCR	Too much template	Dilute template between 10-1000X and redo qPCR to determine template amount that results in a normal qPCR trace
29, 39	Trace shows humps before log-based amplification		qPCR traces will not always look flat at the beginning. Wait, and if traces eventually show normal log-based amplification, no modifications/changes are needed
29, 39	Template only amplifies >40 cycles	Indexing primers degraded	Purchase or use new indexing primers; make a new working indexing primer plate from the stock
33	No PCR bands	PCR requires more cycles	Redo qPCR in Steps 24-29 to determine cycle number
33	No PCR bands	PCR reaction may have been inhibited due to too much template or salt content	Decrease the amount of template added; perform a genomic DNA extraction to get rid of excess salt
49, Supp. Method Step 24	Bands do not run straight or are slightly curved	Too much ethanol	Increase the amount of time beads are allowed to dry in Step 46 or Supp. Method Step 20.
49, Supp. Method Step 24	No bands; beads did not bind to DNA	Nucleic acid buffer was not made correctly; not enough beads	Ensure all reagents are fresh and pH is correct; always make incomplete binding buffer in Supp. Method Step 8 right before use and ensure proportions of reagents are correct; resuspend or mix beads well in Supp. Method Step 1; check that there is no aspiration of beads during washes

2.7 Timing

Temporal recording procedure: ***6 d, 2 h 10 min hands-on***

Transformation of plasmids into expression strain (steps 1-18): **4 d, 1.5 h hands-on.**

Recording transcriptional activity for 48 hours (steps 19-23): **2 d, 40 min hands-on.**

Preparation of CRISPR arrays for deep sequencing and deep sequencing: ***13.5 h, 9.5 h hands-on****, including optional step*

Determine cycle number for first round PCR amplification (steps 24-29): **2 h, 2h hands-on.**

*Note: should only have to be performed once for a given experimental paradig*m

Optional step: Preparation and cleaning of Sera-Mag beads for DNA clean-up (Supplementary Methods): **2 h, 1.5 h hands-on.**

Amplifying and indexing samples (steps 30-49): **6 h, 4 h hands-on.**

Deep sequencing of CRISPR array (steps 50-57): **3.5 h, 2 h hands-on.**

Data analysis: ***1-3 h (depending on number of CPU cores available), 30 min hands-on***

Installing dependencies (steps 58-62): **10 min hands-on.**

Loading experiment sheet; downloading FASTQs from SRA and trimming them (steps 63-64): **30 min, 5 hands-on**.

Extracting new spacers and storing them according to their characteristics and the characteristics of the CRISPR arrays from which they were extracted; storing the data in a data-frame (steps 65-68): **1-3h, 2 min hands-on.**

Determining the order of spacers in each sequenced CRISPR array; storing the data in a data-frame (steps 69-70): **5 min, 1 min hands-on.**

Calculating the ordering scores; storing the data in a data-frame (steps 71-72): **1 min, 1 min hands-on.**

Plotting the data (step 73): **1 min, 1 min hands-on.**

2.8 Anticipated results

Following this protocol, we expect users to be able to generate plots depicting the ordering scores from different transcriptional programs. As an example, we provide the outcome plot using our updated computational pipeline on raw sequencing reads previously obtained[21] from a 48-hour transcriptional recording experiment, where an anhydrotetracycline-induced promoter ("A") was turned on for 24 hours then a choline chloride-induced promoter ("B") was turned on another 24 hours, or vice versa (**Fig. 2.4**).

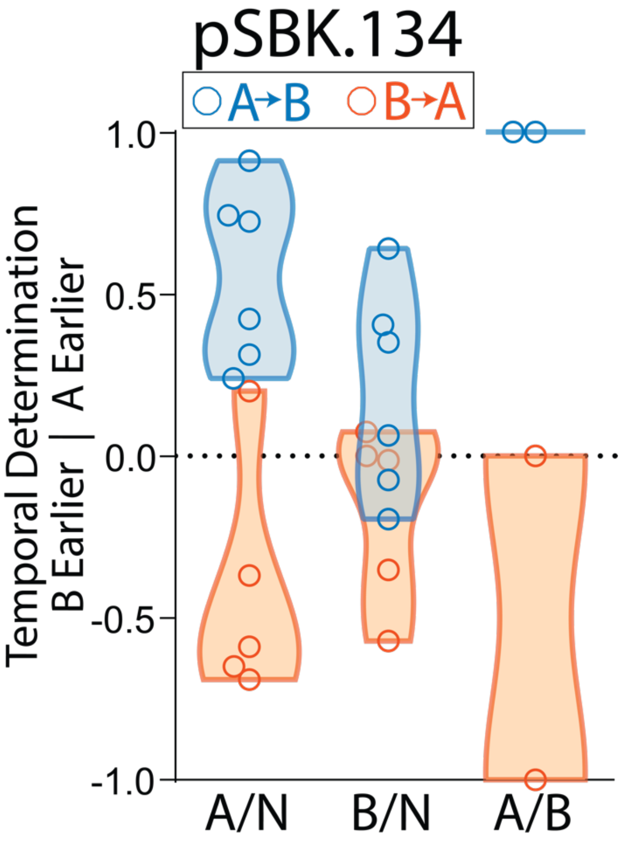


Figure 2.4. Illustrative ordering analysis of a recording experiment. Ordering scores for 48-hour transcriptional recording using sequential 24-hour expression of anhydrotetracycline (“A”) and chlorine chloride (“B”)-induced promoters. When A occurs before B, the calculation of ordering scores results in positive values, suggesting that the analysis pipeline appropriately identified the order of expression where A occurs before B. Likewise, when B occurs before A, the calculation of ordering scores result in negative values, suggesting again that the analysis pipeline appropriately identified the opposite order of expression where B occurs before A. This figure is a re-analysis using our updated computational pipeline on the same raw sequencing reads previously obtained from our previous publication to create Fig. 4L[21]. Open circles correspond to N=6 biological replicates.

Although **Fig. 2.4** depicts a simple transcriptional program, the three ordering scores we describe also enable the representation of more complex transcriptional programs. The only requirement is that each program must be separated into two distinct epochs where the acquisition rate of a given transcriptional signal in each epoch assumed to be constant. A key to the meaning of different ordering score plots is provided (**Fig. 2.2a**), followed by four hypothetical transcriptional programs (**Fig. 2.2b-e**). For each transcriptional program, we simulated the expected ordering score results from six replicates using 2.1 million reads per replicate to give users an intuition of the types of results they can expect for different transcriptional programs. Although we chose programs where transcriptional signal A and B each have the same acquisition rate, the exact magnitude of ordering scores can also reflect differences between the strengths and resulting varied acquisition rates of different signals as well.

We anticipate that users could perform an experiment akin to our 48 h recordings, run the ordering score analysis, and plot them. By comparing the distribution of their scores with the key in **Fig. 2.2a**, as well as the different possibilities illustrated in **Fig 2.2b-e**, and together with the interactive simulations provided in the accompanying notebook, we believe that users will be able to make inferences regarding the underlying transcriptional programs that took place during the recording experiment.

2.9 Supplemental Methods

2.9.1 Supplemental Materials

2.9.1.1 Reagents

Sera-Mag bead preparation and testing

- Sera-Mag^{TM} Magnetic SpeedBeads, carboxylated, 1 um, 3 EDAC/PA5 (GE Healthcare Life Sciences #65152105050250) !CAUTION contains 0.05% sodium azide, which is toxic; avoid contact with skin or eyes.

- HCl solution,1.0 N (Sigma-Aldrich, cat. no. H9892-100ML) !CAUTION HCl is corrosive and an irritant; avoid contact with skin and eyes. When handling HCl, wear gloves and eye protection.
- Tris base (1M, add 6.057 g in 50 mL UltraPure^{TM} Distilled Water, sterile-filtered; Thermo Fisher Scientific, cat. no. BP152-500)
- NaCl (5M, add 14.610 g in 50 mL UltraPure^{TM} Distilled Water; Thermo Fisher Scientific, cat. no. S271-3)
- Disodium-EDTA (0.1M, add 1.816 g in 50 mL UltraPure^{TM} Distilled Water; J.T. Baker, cat. no. 6381-92-6) !CAUTION Disodium-EDTA is toxic if swallowed and is an irritant to skin and eyes. When handling disodium-EDTA, wear gloves and eye protection.
- Tween 20, non-ionic, aqueous solution, 10% (w/v) (Sigma-Aldrich, cat. no. 11332465001)
- PEG 8000 (Sigma-Aldrich, cat. no. 89510-1KG-F)

Equipment

- LP Vortex Mixer Thermo Fisher Scientific, cat. no. 88880017)
- Magnetic rack (MagRack 6) for 1.5 mL microcentrifuge tubes bead magnet (GE Healthcare Life Sciences, cat. no. 26980)

2.9.1.2 Reagent Setup

50%(w/v) PEG 8000

Add 12.5 g PEG 8000 in 14 mL UltraPure^{TM} Distilled Water. Shake to dissolve and let incubate on benchtop at room temperature for at least 1 hour until all bubbles dissipate. Add distilled water until solution reaches a total volume of 25 mL. Mix well. Store at 4°C for up to 1 year.

Tween-TE DNA Binding Buffer

Mix 48.564 mL UltraPure^{TM} Distilled Water, 0.5 mL 1M Tris base, 0.5 mL 0.1M disodium-EDTA, 0.25 mL 10% (v/v) Tween 20, and 0.186 mL 1N HCl. Make fresh, right before use.

Nucleic acid incomplete binding buffer

Mix 25 mL 5M NaCl, 3.582 mL UltraPure^{TM} Distilled Water, 0.5 mL 1M Tris base, 0.5 mL 0.1M Disodium-EDTA, and 0.168 mL 1N HCl.

CRITICAL Prepare nucleic acid incomplete binding buffer fresh during bead washing steps (see Supplementary Methods step 8).

2.9.2 Supplemental Protocol

2.9.2.1 Preparation of CRISPR arrays for deep sequencing: Preparation and cleaning of Sera-Mag beads for DNA clean-up

● Timing 2 h, 1.5 h hands-on

CRITICAL This supplementary protocol prepares Sera-Mag beads for DNA clean-up, which will be used in steps 41-48 of the main protocol. To save time, users should also consider purchasing AMPure XP beads instead, which can be immediately used without any additional preparation or clean-up. AMPure XP beads, however, are very expensive, so we offer this alternative protocol to reduce the cost of performing the protocol.

1. *Clean Sera-Mag beads.* This step cleans Sera-Mag beads and stores them in a nucleic acid binding buffer used during PCR clean-up. Once an aliquot of beads has been cleaned and moved into binding buffer, it can be reused for several months if properly stored. Invert the bottle containing Sera-Mag beads and pipette the solution up and down several times to resuspend beads well.
2. Transfer 1 mL of resuspended beads to a 1.5 mL microcentrifuge tube.
3. Place the microcentrifuge tube on magnet until beads migrate near the magnet and the solution is clear (~30 s).
4. Remove the supernatant.
5. Add 1 mL of DNA buffer (see TE-Tween DNA Buffer in Reagent Setup) to the bead pellet and close the microcentrifuge tube.
6. Remove the microcentrifuge tube from the magnet and resuspend beads by vortexing for at least 15 s. Following mixing, the solution should appear cloudy and homogenous. Spin down the liquid using a microcentrifuge.
7. Repeat Supplementary Method steps 3-6 twice more for a total of 3 washes.
8. Prepare 29.75 mL of freshly made nucleic acid incomplete binding buffer (see Reagent Setup) in a 50 mL conical tube.
9. Remove supernatant from beads and immediately add 1 mL of incomplete binding buffer into the microcentrifuge tube while still on the magnet.
10. Remove the microcentrifuge tube from the magnet and resuspend beads by vortexing for at least 15 s. If liquid is stuck onto the sides, briefly spin down the microcentrifuge tube in a microcentrifuge but be careful not to also pellet the beads.

11. Transfer the 1 mL of beads in the incomplete binding buffer to the conical tube containing the rest of the incomplete binding buffer. Cap the tube and vortex for at least 30 s until beads are well mixed into the entire buffer.
12. Using a 25 mL serological pipette, add 20 mL of 50% (w/v) PEG stock (see Reagent setup) to the conical tube. Dispense slowly to allow the viscous liquid to slide down the inside walls of the pipette to ensure an accurate volume of 50% PEG is added.
13. Add 0.25 mL 10% (w/v) Tween 20.
14. Cap the tube and mix solution through inversion gently until the color of the solution appears homogenous. This solution contains prepared and cleaned beads in complete binding solution that can be immediately used for DNA clean-up. To distinguish between Sera-Mag beads that still need to be prepared and those in complete binding solution that are ready to be used for DNA clean-up, these prepared beads will be referred to as SPRI beads.

    PAUSE POINT SPRI beads can be stored at 4°C for at least 1 year.

    CRITICAL STEP Prior to each use, ensure solution has come to room temperature and is thoroughly mixed through inversion as beads will pellet at the bottom of the tube over time.

15. *Testing SPRI bead ratio for DNA clean-up.* This step determine what ratio of beads to DNA should be used to clean-up DNA without selecting for size. Mix 16 µL 1 kB+ DNA ladder with 144 µL water in a microcentrifuge tube. Pipette 20 µL of diluted ladder each into eight clean, new microcentrifuge tubes
16. Create dilution series of 8 different bead-to-ladder ratios in the microcentrifuge tubes each containing 20 µL diluted ladder according to **Table 2.12**:

Table 2.12 Dilution series for bead:DNA mixtures

Microcentrifuge tube	Amount of beads per tube from step 14 of Supp. Method (μL)	Final bead-to-DNA dilution
1	10	0.5X
2	14	0.7X
3	18	0.9X
4	22	1.1X
5	26	1.3X
6	30	1.5X
7	34	1.7X
8	38	1.9X

17. Incubate the dilutions for 5 min at room temperature. Place all tubes on a magnet until the beads migrate and the solution clears (~1 min). Remove supernatant.
18. Add 200 μL fresh 70% ethanol to each tube and let incubate on the magnet until solution clears (~1 min). Remove supernatant.
19. Repeat step Supplemental Method step 18 for an additional wash.
20. Let beads dry for <3 minutes.

    CRITICAL STEP DNA needs to dry to remove contaminant ethanol but allowing the beads to dry for too long will result in low recovery. The presence of cracks appearing in the beads is a sign that the DNA has been allowed to dry for too long.
21. Remove the plate from the magnet and resuspend the DNA with 25 μL water. Pipette up and down between 10-15 times to mix. Incubate at 5 minutes at room temperature.
22. Place the tubes back onto the magnet until the solution clears. Collect supernatant and move into new, fresh microcentrifuge tube.

23. Run the dilution series on a 2% agarose E-gel EX by adding 20 μL of each dilution per well. Also include the 1 kB+ ladder in the gel by adding 2 μL undiluted ladder and 18 μL water to the marker lane. Run gel for 10 minutes.

24. Determine the ideal ratio to use by selecting the smallest ratio that still retains most of the smaller size DNA band. An example gel of different dilutions ratios run along with a ladder is shown below (Supplemental Figure 1). Based on this gel, a bead-to-DNA ratio of 1.5X would be selected because it is the smallest ratio that nonetheless retains the 100-nucleotide long DNA band. Once a ratio has been determined for a given aliquot of beads, there is typically no need for re-testing.

    CRITICAL STEP a typical bead-to-DNA ratio in our hands is between 1.3X – 2X.

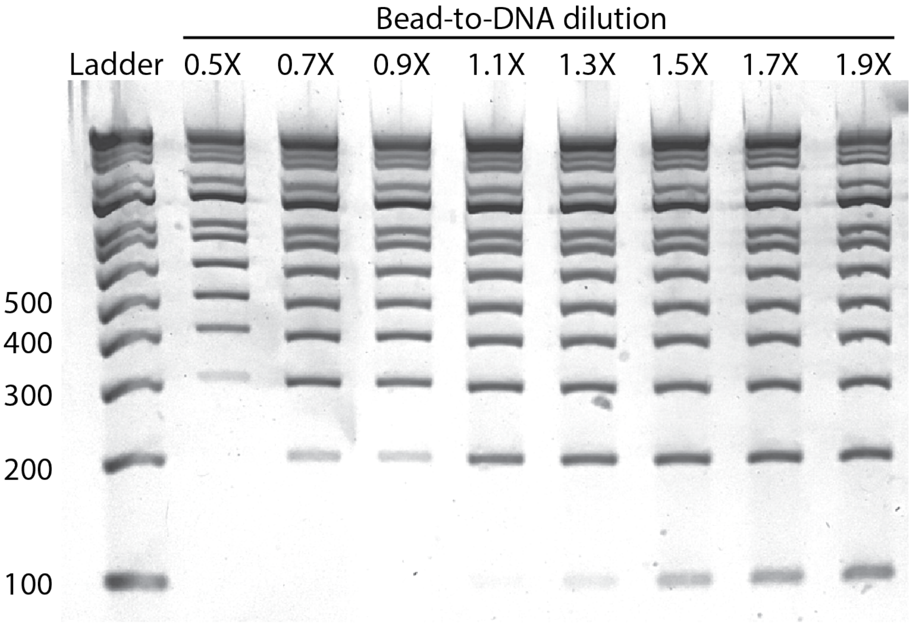


Figure 2.5. Representative example of gel used to test bead-to-DNA ratios.

2.10 References

1. Siuti, P., Yazbek, J. & Lu, T. K. Synthetic circuits integrating logic and memory in living cells. *Nat. Biotechnol.* **31**, 448–452 (2013).
2. Bonnet, J., Yin, P., Ortiz, M. E., Subsoontorn, P. & Endy, D. Amplifying Genetic Logic Gates. *Science* **340**, 599–603 (2013).
3. Yang, L. *et al.* Permanent genetic memory with >1-byte capacity. *Nat. Methods* **11**, 1261–1266 (2014).
4. Farzadfard, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272 (2014).
5. Courbet, A., Endy, D., Renard, E., Molina, F. & Bonnet, J. Detection of pathological biomarkers in human clinical samples via amplifying genetic switches and logic gates. *Sci. Transl. Med.* **7**, 289ra83-289ra83 (2015).
6. Roquet, N., Soleimany, A. P., Ferris, A. C., Aaronson, S. & Lu, T. K. Synthetic recombinase-based state machines in living cells. *Science* **353**, aad8559 (2016).
7. Hsiao, V., Hori, Y., Rothermund, P. W. & Murray, M. M. A population-based temporal logic gate for timing and recording chemical events. *Mol. Syst. Biol.* **12**, 869 (2016).
8. Weinberg, B. H. *et al.* Large-scale design of robust genetic circuits with multiple inputs and outputs for mammalian cells. *Nat. Biotechnol.* **35**, 453–462 (2017).
9. Perli, S. D., Cui, C. H. & Lu, T. K. Continuous genetic recording with self-targeting CRISPR-Cas in human cells. *Science* **353**, aag0511 (2016).
10. Tang, W. & Liu, D. R. Rewritable multi-event analog recording in bacterial and mammalian cells. *Science* **360**, eaap8992 (2018).

11. Kempton, H. R., Love, K. S., Guo, L. Y. & Qi, L. S. Scalable biological signal recording in mammalian cells using Cas12a base editors. *Nat. Chem. Biol.* 1–9 (2022) doi:10.1038/s41589-022-01034-2.

12. Chen, W. *et al*. Multiplex genomic recording of enhancer and signal transduction activity in mammalian cells. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.05.467434 (2021).

13. Loveless, T. B. et al. Molecular recording of sequential cellular events into DNA. Preprint at *bioRxiv* https://doi.org/10.1101/2021.11.05.467507 (2021).

14. Choi, J. *et al.* A time-resolved, multi-symbol molecular recorder via sequential genome editing. *Nature* **608**, 98–107 (2022).

15. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. Molecular recordings by directed CRISPR spacer acquisition. *Science* **353**, aaf1175 (2016).

16. Sheth, R. U., Yim, S. S., Wu, F. L. & Wang, H. H. Multiplex recording of cellular events over time on CRISPR biological tape. *Science* **358**, 1457–1461 (2017).

17. Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380–385 (2018).

18. Yim, S. S. *et al.* Robust direct digital-to-biological data storage in living cells. *Nat. Chem. Biol.* **17**, 246–253 (2021).

19. Sheth, R. U. & Wang, H. H. DNA-based memory devices for recording cellular events. *Nat. Rev. Genet.* **19**, 718–732 (2018).

20. Lear, S. K. & Shipman, S. L. Molecular recording: transcriptional data collection into the genome. *Curr. Opin. Biotechnol.* **79**, 102855 (2023).

21. Bhattarai-Kline, S. *et al.* Recording gene expression order in DNA by CRISPR addition of retron barcodes. *Nature* (2022) doi:10.1038/s41586-022-04994-6.

22. Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.* **40**, 5569–5576 (2012).

23. Nuñez, J. K. *et al.* Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nat. Struct. Mol. Biol.* **21**, 528–534 (2014).

24. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345–349 (2017).

25. Silas, S. *et al.* Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase–Cas1 fusion protein. *Science* **351**, aad4234 (2016).

26. Tanna, T., Schmidt, F., Cherepkova, M. Y., Okoniewski, M. & Platt, R. J. Recording transcriptional histories using Record-seq. *Nat. Protoc.* 1–27 (2020) doi:10.1038/s41596-019-0253-4.

27. Schmidt, F. *et al.* Noninvasive assessment of gut function using transcriptional recording sentinel cells. *Science* **376**, eabm6038 (2022).

28. Yehl, K. & Lu, T. Scaling computation and memory in living cells. *Curr. Opin. Biomed. Eng.* **4**, 143–151 (2017).

29. Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol. Cell* **62**, 824–833 (2016).

30. Yoganand, K. N. R., Sivathanu, R., Nimkar, S. & Anand, B. Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. *Nucleic Acids Res.* **45**, 367–381 (2017).

31. Sharon, E. *et al.* Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* **175**, 544-557.e16 (2018).

32. Kong, X. *et al.* Precise genome editing without exogenous donor DNA via retron editing system in human cells. *Protein Cell* **12**, 899–902 (2021).

33. Lopez, S. C., Crawford, K. D., Lear, S. K., Bhattarai-Kline, S. & Shipman, S. L. Precise genome editing across kingdoms of life using retron-derived DNA. *Nat. Chem. Biol.* **18**, 199–206 (2022).

34. Zhao, B., Chen, S.-A. A., Lee, J. & Fraser, H. B. Bacterial Retrons Enable Precise Gene Editing in Human Cells. *CRISPR J.* **5**, 31–39 (2022).

35. Palka, C., Fishman, C. B., Bhattarai-Kline, S., Myers, S. A. & Shipman, S. L. Retron reverse transcriptase termination and phage defense are dependent on host RNase H1. *Nucleic Acids Res.* **50**, 3490–3504 (2022).

36. Munck, C., Sheth, R. U., Freedberg, D. E. & Wang, H. H. Recording mobile DNA in the gut microbiota using an Escherichia coli CRISPR-Cas spacer acquisition platform. *Nat. Commun.* **11**, 95 (2020).

37. Lee, P.Y., Costumbrado, J., Hsu, C.Y., Kim, Y.H. Agarose Gel Electrophoresis for the Separation of DNA Fragments. J. Vis. Exp. (62), e3923, doi:10.3791/3923 (2012).

38. Kluyver, T., *et al.* Jupyter Notebooks—a publishing format for reproducible computational workflows. In: Loizides, F. & Schmidt, B. (eds.) *Positioning and Power in Academic Publishing: Players, Agents and Agendas,* 87-90 (IOS Press, 2016).

39. Joshi, N.A. & Fass, J.N. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) Available at: https://github.com/najoshi/sickle. (2011).

40. JoVE Science Education Database. *Microbiology.* Serial Dilutions and Plating: Microbial Enumeration. JoVE, Cambridge, MA, (2022).

Chapter 3: Enabling *in vivo* production of DNA donor templates in mammalian cells

3.1 Introduction

DNA donor templates are commonly used to precisely edit a cell's genome by using homology-directed repair (HDR), which uses a template to alter or insert new genetic sequences into the genome following a double-stranded break (DSB) created by a nuclease such as Cas9. While this template is typically synthesized *ex vivo*, delivering this template in high abundance along with a nuclease into cells is a great technical challenge that may contribute to low rates of precise editing and unintended insertions or deletions[1–4].

One way to overcome this challenge is to produce DNA template *in vivo* by expressing a retron reverse transcriptase that can specifically reverse transcribe a sequence attached to guide RNA (gRNA) into a DNA template. This strategy has previously been used to create edits in bacteria[5,6] and yeast[7]. In this chapter, I will highlight pipelines I developed to enable *in vivo* production of DNA donor template, and resulting precise editing, in cultured mammalian cells. As a proof-of-concept, mammalian retron-based editing was first performed by delivering plasmids that expressed the necessary gRNA. Afterwards, retron editing was then performed by delivering RNA instead of plasmids, which is ultimately more relevant to current therapeutic delivery approaches in higher-order mammalian models and also allows us to test the insertion of longer exons.

3.2 Plasmid-based delivery

We first sought to test whether retron-produced DNA could be used for precise editing of cultured mammalian cells by delivering the necessary retron non-coding RNA (ncRNA) and gRNA on a single plasmid into HEK293T cells engineered to express both Cas9 and the retron

Eco1 reverse transcriptase (Eco1 RT). To simplify the genetic construct inserted into human cells as compared to yeast, where Cas9 and Eco1 RT were expressed separately, a colleague Santi Lopez tested six different single-promoter architectures in yeast: four fusion proteins with both orientations of Cas9 and Eco1 RT using two different linker sequences, and two versions where Cas9 and Eco1 RT were separated by a P2A sequence[8] in both possible orientations. Out of all architectures tested, Lopez found that the Cas9-P2A-RT version resulted in the highest ADE2 editing rates (**Figure 3.1a**). As a result, a cassette containing the best-performing Cas9-P2A-RT architecture under a doxycycline-inducible promoter was integrated into a wild-type HEK293T line using a Piggybac transposon system. As a negative control, a cassette containing only Cas9 was also integrated into another HEK293T cell line.

Afterwards, a plasmid containing a designed retron ncRNA/gRNA driven by a polymerase III H1 promoter was transiently transfected either of the two engineering HEK293T cell lines (**Figure 3.1b**). In each case, the plasmid was designed to target and edit one of six different loci: *HEK2, RNF2, EMX1, FANCF, HEK4*[9]*, or AAVS*1[10] (**Figure 3.1c-h**). To compare editing rates across sites, either HEK293T cell line was induced to express Cas9 and/or Eco1 RT for 24 hours prior to transfecting the ncRNA/gRNA plasmid. Cells were collected at both 1 and 3 days post-transfection, although we found that only cells collected after 3 days contained noticeable editing. Using Illumina sequencing, we found precise editing at a rate of 0.1-2.5% in the presence of the Eco1 RT as compared to a background rate of 0.2% or lower in the absence of Eco1 RT.

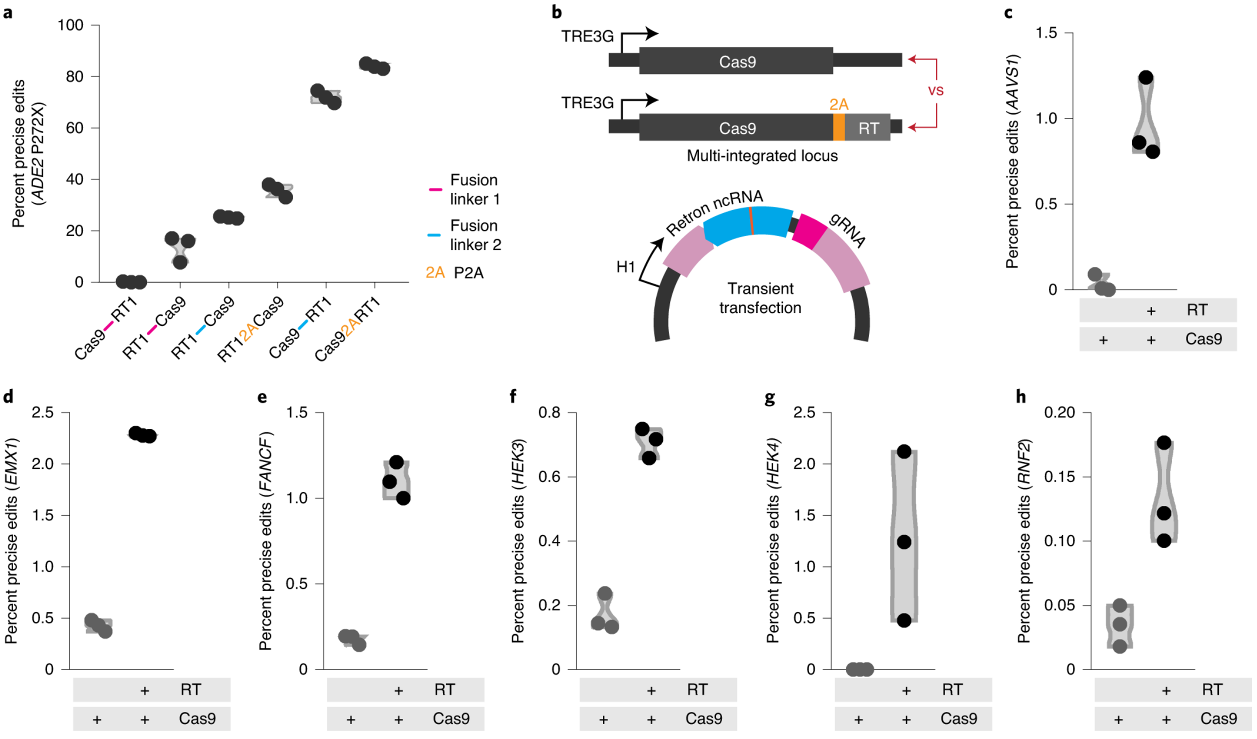


Figure 3.1 Plasmid-based delivery of retron ncRNA/gRNA into mammalian cells to achieve precise editing. (a) Six different single-promoter architectures for editing the ADE2 locus in S. cerevisiae were tested. Fusion linker 1 amino acid sequence is GGTSSGGSGTAGSSGATSGG; fusion linker 2 sequence is SGGSSGGSSGSETPGTSESATPESSGGSSGGSS. Circles show each of the three biological replicates; one-way ANOVA, effect of construct: P <0.0001; n = 3 biological replicates. **(b)** Schematic showing the elements for editing in human cells. Top, integrated protein cassettes that are compared to each other in c-h. Bottom, plasmid for transient transfection of the site-specific ncRNA/gRNA. **(c-h)** Quantification of precise editing of six different loci in HEK293T cells by Illumina sequencing. Circles represent each of the three biological replicates; unpaired t-test: effect of Cas9 alone versus Cas9 and RT. **(c)** *AAVS1* locus. P = 0.0026. **(d)** *EMX1* locus. P < 0.0001. **(e)** *FANCF* locus. P = 0.0001. **(f)** *HEK3* locus. P = 0.0002. **(g)** *HEK4* locus. P = 0.0543. **(h)** *RNF2* locus. P = 0.0158.

3.2 RNA-based delivery

Although we found that retron editing can achieve precise edits in mammalian cells using Lipofectamine-based delivery of plasmids encoding ncRNA/gRNA, such a delivery method would be infeasible in more therapeutically relevant cell lines or live animal models. Furthermore, the use of the polymerase III H1 promoter—which is commonly used to make shorter transcripts—to drive the transcription of ncRNA/gRNA may restrict the length of insertion made.

As a result, we first tested if we could successfully create DNA templates, or RT-DNA, by delivering in-vitro transcribed RNA, rather than plasmids, into HEK293T cells containing a cassette that inducibly expressed a retron reverse transcriptase. To maximize the chances of successfully seeing reverse transcription, I used retron reverse transcriptase Eco2, which the Shipman lab has previously found produces more RT-DNA than Eco1 RT.

RT-DNA production was measured using a slightly modified version of a qPCR-based strategy previously published[11]. I first validated that the original qPCR-based strategy, which compared the amount of target DNA to a sequence on the plasmid expressing the ncRNA, successfully replicated when using primers against different targets and the Eco2 RT (**Figure 3.2a**). I either induced or did not induce expression of a modified HEK293T cell line with an integrated cassette containing both Eco2 RT and wild-type Eco2 ncRNA using doxycycline. One day later, DNA was harvested from both the induced and uninduced cell lines. I then examined DNA levels of two different targets, Eco2 RT-DNA and human mitochondrial COX2 gene (hMT-CO2)—whose abundance should be unaffected by doxycycline and Eco2 RT—to the control gene RPL13A1 using the deltadeltaCt method. As expected, induction of the Eco2 RT resulted in a 10 to 30-fold-increase in RT-DNA as compared to no induction. In comparison, induction of Eco2 RT had no effect on the negative control target hMT-CO2, suggesting that the qPCR assay and design was sensitive enough to detect the production of Eco2 RT-DNA.

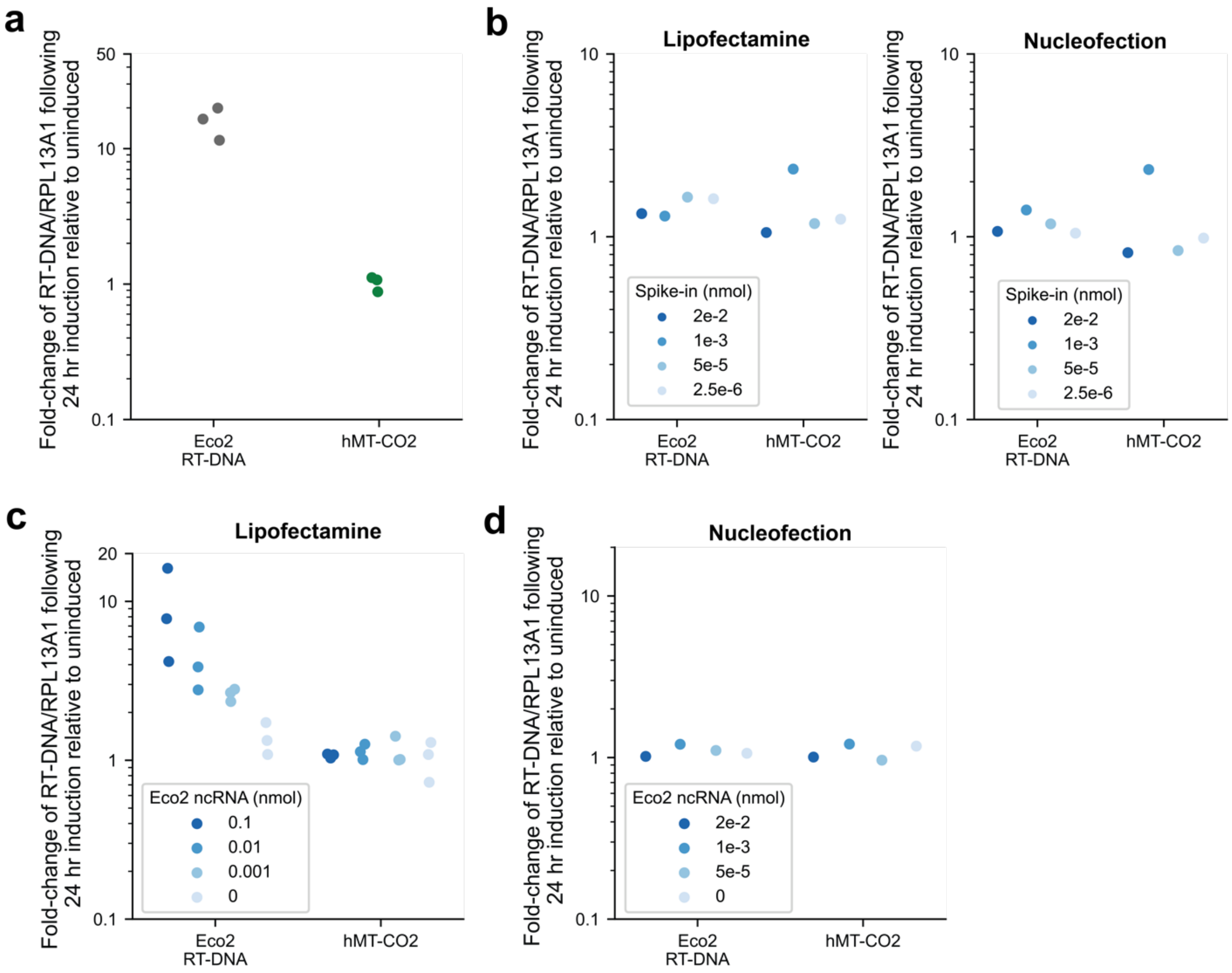


Figure 3.2 Developing an assay to measure Eco2 RT-DNA abundance using RNA delivery. **(a)** Fold-change in amount of either Eco2 RT-DNA or hMT-CO2 DNA when both Eco2 RT and Eco2 ncRNA is expressed in HEK293T cells relative to when they are not expressed. Circles show each of the three biological replicates. **(b)** Fold-change in amount of either Eco2 RT-DNA or hMT-CO2 DNA when Eco2 RT is induced relative to non-induced following delivery of different concentrations of DNA reference "spike-in." Left, spike-in delivery using Lipofectamine 3000. Right, spike-in delivery using nucleofection. Circles show each of one biological replicate. **(c)** Fold-change in amount of either Eco2 RT-DNA or hMT-CO2 DNA when Eco2 RT is induced relative to non-induced following Lipofectamine-based delivery of spike-in and different concentrations of Eco2 ncRNA. Circles show each of the three biological replicates. **(d)** Fold-change in amount of either Eco2 RT-DNA or hMT-CO2 DNA when Eco2 RT is induced relative to non-induced following nucleofection-based delivery of spike-in and different concentrations of Eco2 ncRNA. Circles show each of one biological replicate.

Given that the qPCR strategy was validated, I next turned to using the same strategy to determine the abundance of RT-DNA when delivering RNA rather than plasmids. However, since there would no longer be a plasmid sequence to act as a reference, a new control would be required for the deltadeltaCt method. As a result, I synthesized a DNA spike-in whose sequence would be differentiated from the wild-type RT-DNA by five additional nucleotide bases and could be co-delivered along with the Eco2 ncRNA. To determine the proper concentration of DNA spike-in to use for the modified qPCR strategy, I used either Lipofectamine 3000 or nucleofection to deliver DNA spike-in into a HEK293T cell line containing an integrated cassette with an inducible Eco2 RT at different concentrations without any RNA (**Figure 3.2b**). I found that, regardless of the amount of DNA spike-in delivered, the qPCR was accurately determine that the amount of DNA spike-in was unaffected regardless of if the cells expressed Eco2 RT. Moving forward, a spike-in of 5x10^{-5} nmol for a nucleofection-based strategy and a spike-in of 2.5x10^{-6} nmol for a Lipofectamine-based strategy was used when co-delivering ncRNA, as these specific values were the lowest that were still detectable by the qPCR machine.

After validating the modified qPCR assay when using RNA delivery, the amount of RT-DNA produced in induced relative to non-induced cells was quantified after delivering different titrations of ncRNA using Lipofectamine (**Figure 3.2c**) or nucleofection (**Figure 3.2d**). We found the delivery of RNA using Lipofectamine resulted in clear production of RT-DNA whose levels covaried directly relative to the amount of RNA delivered. In comparison, no clear evidence exists that RT-DNA was produced post-nucleofection, which I hypothesize may be due to the RNA being degraded too quickly prior to reverse transcription.

Following validation that RT-DNA can be produced using RNA delivery, I next attempted to show that we can achieve retron editing using RNA delivery. The ncRNA-gRNA to induce a single nucleotide base change and modification of the PAM site in the *EMX1* site was synthesized, chosen because that site resulted in the highest rates of editing when delivered using plasmids (**Figure 3.1d**). RNA was delivered into an already induced HEK293T cell line containing an integrated cassette with both Cas9 and Eco1 RT, and rates of precise editing were evaluated at 24- and 48-hours post-transfection using Illumina-based sequencing for a variety of different RNA titrations (**Figure 3.3**). In addition, both Lipofectamine and nucleofection-based strategies were used to deliver different amounts of ncRNA-gRNA, although the amount of DNA delivered using nucleofection was increased compared to the amounts used for measuring RT-DNA production since we originally saw no RT-DNA production (**Figure 3.2d**). In contrast to our previous findings, Lipofectamine-based delivery resulted in almost undetectable rates of editing regardless of the amount of RNA used. Meanwhile, when the maximum possible amount of RNA was delivered using nucleofection, precise editing reliably occurred at a rate between 0.6-2%.

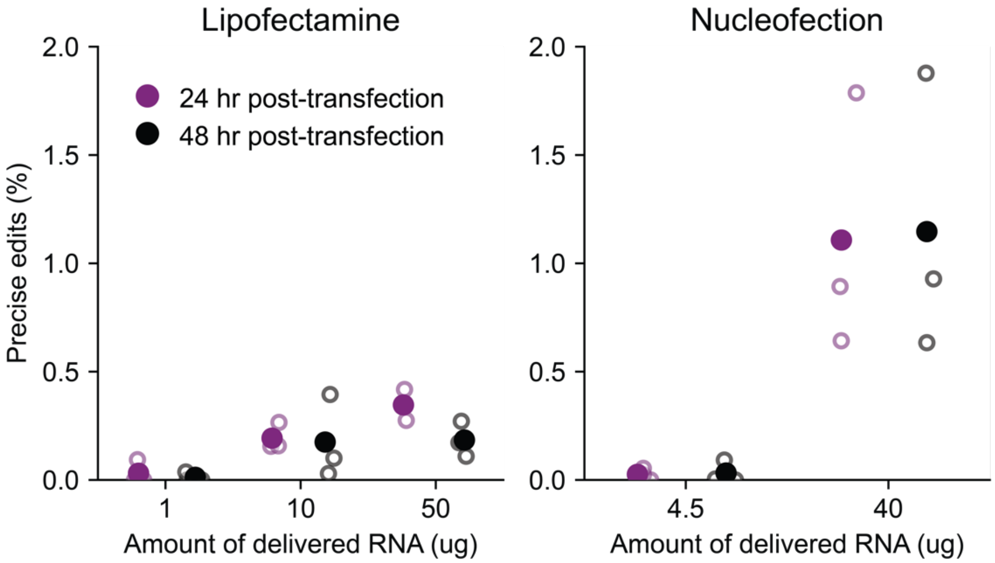


Figure 3.3 Nucleofection-based delivery of retron ncRNA/gRNA into mammalian cells to achieve precise editing. Percentage of precise edits seen in the EMX1 site either 24 or 48 hours after delivery of ncRNA-gRNA into HEK293T cells expressing Cas9 and Eco1 RT. Left, rates seen when delivering RNA via Lipofectamine 3000. Right, rates seen when delivering RNA via nucleofection. Open circles represent each of 3 biological replicates and closed circles represent the mean.

Achieving editing following delivery of a ncRNA-gRNA is promising, but the greatest translational benefit requires all components of the retron editor, including Cas9 and the retron RT itself, to also be delivered as RNA. However, despite trying to deliver Cas9 mRNA, Eco1 RT mRNA, and ncRNA-gRNA at different ratios and amounts using nucleofection, we were unable to detect any precise editing.

Although more optimization is necessary before retron editing can be performed purely through RNA delivery, the ability to perform retron-based editing using ncRNA-gRNA exogenously synthesized prior to delivery nonetheless allows us to explore certain aspects of retron editing that would otherwise be constrained when using a polymerase III H1 promoter to generate the ncRNA-gRNA *in vivo*. Specifically, polymerase III promoters are generally used to create small RNAs and are not ideal for creating larger insertions. As a result, we tested if the retron was processive enough to insert longer insertions by delivering a ncRNA-gRNA containing either a ~25-nucleotide FLAG tag, or a ~1 kB GFP exon (**Figure 3.4**). We found that we were able to detect precise insertions of both FLAG tag and GFP by performing a PCR where one primer sits right outside the genomic insertion site and another primer sits on the specific insertion site itself (**Figure 3.4a, b**). Furthermore, we could quantify editing rate of FLAG tag insertion using Illumina-based sequencing (**Figure 3.4c**). While insertions can be detected regardless of size, we found that editing rate drops as the length of the insertion grows. Indeed, FLAG tag was only inserted at a rate of around 0.25% as compared to a base transversion.

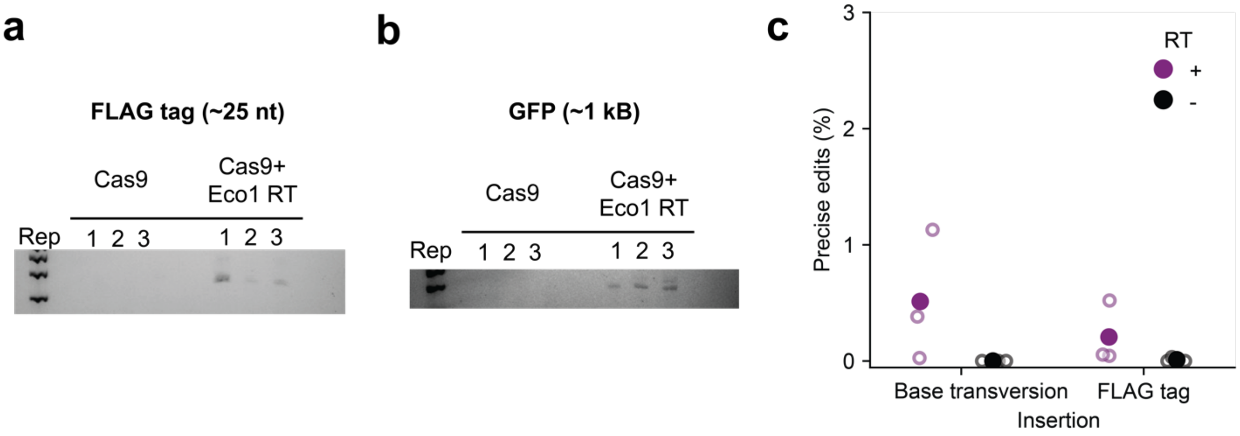


Figure 3.4 Retron editing enables exon-sized insertions in mammalian cells. (a) Gel showing presence or absence of FLAG tag in samples expressing Cas9 with or without Eco1 RT after nucleofection of ncRNA-gRNA with a FLAG tag insertion. Primers were used that bind immediately outside the insertion site as well as complementary to the insertion sequence itself, so presence of a band suggests that an insertion occurred. **(b)** Gel showing presence or absence of FLAG tag in samples expressing Cas9 with or without Eco1 RT after nucleofection of ncRNA-gRNA with a GFP gene insertion. Primers were used that bind immediately outside the insertion site as well as complementary to the insertion sequence itself, so presence of a band suggests that an insertion occurred. **(c)** Quantification of editing efficiency of either a base transversion or insertion of a FLAG tag using Illumina-based sequencing.

The presence of exon-sized edits does suggest that retron editing is capable of inserting long sequences, although further optimization is necessary to improve upon the current low editing rates. One avenue to improve editing rates is to screen other retron RTs to determine if others may result in higher editing efficiencies. As a proof-of-concept, two other retron RTs—Eco3 and Eco5—were used to modify the EMX1 site through the insertion of a 10-nt barcode (**Figure 3.5**). Eco3 RT outperformed Eco1 RT even when recruited to perform longer and more complex edit, suggesting that screening additional retron RTs may be a suitable way to improve the efficiency of retron editors.

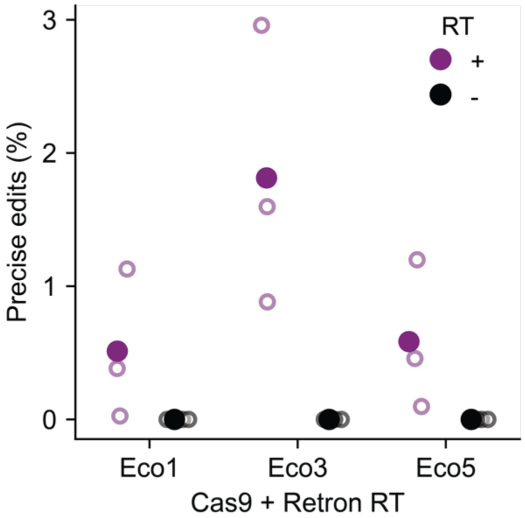


Figure 3.5 Comparing editing rates across three different retron RTs. The precise editing rate of a single base change for Eco1 RT was compared to the rate of precise insertion of a 10-nt barcodes when using cell lines where Eco1 RT was replaced with either Eco3 RT or Eco5 RT.

3.3 Discussion

In this chapter, we show as a proof-of-principle that retron editing can produce precise edits in cultured human cells using both plasmid- and RNA-based delivery strategies. Furthermore, the processivity of the retron RT allows it to even perform exon-sized insertions, a clear advance beyond similar technology prime editing, whose editing rate falters after 40 base pairs[9]. Although these experiments demonstrate the promise of retron-based editing, efficiency in human cells is currently low and needs to be improved before the technology can be used in future therapeutic applications. However, we also show a potential path towards improvement by showing encouraging evidence that editing efficiencies may be improved by further screening and testing alternative retron RTs.

3.4 References

1. Luo, D. & Saltzman, W. M. Synthetic DNA delivery systems. *Nat. Biotechnol.* **18**, 33–37 (2000).
2. Lin, S., Staahl, B. T., Alla, R. K. & Doudna, J. A. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* **3**, e04766 (2014).
3. Paquet, D. *et al.* Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9. *Nature* **533**, 125–129 (2016).
4. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, 765–771 (2018).
5. Farzadfard, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272 (2014).
6. Schubert, M. G. *et al.* High-throughput functional variant screens via in vivo production of single-stranded DNA. *Proc. Natl. Acad. Sci.* **118**, e2018181118 (2021).
7. Sharon, E. *et al.* Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* **175**, 544-557.e16 (2018).
8. Liu, J.-J. *et al.* CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* **566**, 218–223 (2019).
9. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* (2019) doi:10.1038/s41586-019-1711-4.
10. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339**, 823–826 (2013).

11. Lopez, S. C., Crawford, K. D., Lear, S. K., Bhattarai-Kline, S. & Shipman, S. L. Precise genome editing across kingdoms of life using retron-derived DNA. *Nat. Chem. Biol.* **18**, 199–206 (2022).

Chapter 4: Importing DNA writers into mammalian mitochondria

4.1 Introduction

Synthetic biologists increasingly leverage natural mitochondrial protein import pathways for compartmentalized metabolic engineering[1,2] and the development of molecular therapeutics[3,4]. For metabolic engineering, sequestering enzymes within yeast mitochondria has resulted in a ~300-fold increase in production of high-value biosynthetic compounds[5,6]. In human health, emerging mitochondrial therapeutics address a major unmet need since mutations to mitochondrial DNA (mtDNA) are at the root of numerous incurable diseases that affect over 1 in 5000 individuals[7,8]. In search of cures, researchers are exploring both allotopic expression, where corrected mitochondrial genes are expressed from the nuclear genome and sent to the mitochondria[9], and gene editing, where mutated mtDNA is either depleted by nucleases or corrected by base editors[10–18].

All of these approaches require efficient targeting of proteins of interest (POIs) to the mitochondria. The most common strategy to achieve such localization is by fusing a mitochondrial targeting sequence (MTS), typically a short and positively charged signal peptide, to the N-terminus of the POI. MTSs are recognized by translocases on the outer and inner mitochondrial membrane (TOM/TIM23 complex) that import the POI through both mitochondrial membranes and release the protein into the mitochondrial matrix following cleavage of the N-terminus MTS from the POI[19,20]. Hundreds of putative MTSs have been identified from natural proteins using computational tools[21–24].

However, attachment of an individual MTS to a given POI does not always guarantee efficient import into mitochondria[25,3]. In fact, allotopic expression and gene editing approaches in mammalian mitochondria have been hindered by low or non-specific POI localization. For instance, only a small sub-selection of protein subunits typically encoded in mtDNA are able to be allotopically expressed in mammalian cells[26,27]. Moreover, even when proteins are imported to mammalian mitochondria, they can also accumulate in other organelles[28]. For mitochondrial gene editing, such imprecise localization poses danger; for example, the mitochondrial base editor DdCBE[16] has substantial off-target editing in nuclear DNA[29], highlighting the need for more specific import of genome editing-related proteins to mitochondria.

Previous studies in yeast suggest that length, hydrophobicity, charge, and folding of the MTS or POI can all affect the efficiency of mitochondrial import[30–32,19]. However, research in mammalian cells is much sparser and at present a given MTS-POI combination cannot be assumed to result in reliable mitochondrial import. Instead, researchers often empirically test multiple MTSs before finding one that results in their specific POI localizing in mitochondria[33,15].

To address a relative lack of broad experimental data and help establish a quantitative assessment of mitochondrial localization in mammalian cells, we developed a quantitative and high-throughput imaging-based pipeline to measure POI import into mitochondria. Using this platform, we screened combinations of three commonly used N-terminus MTSs and POIs from five protein families relevant to mitochondrial gene editing to reveal the most reliable MTS-POI combinations.

4.2 Results

4.2.1 High-throughput Localization Workflow

To investigate the effect of MTS on mitochondrial import across different POIs, we generated 66 protein cassettes containing a combination of localization signal and POI, followed by a HA tag on the C-terminus (**Fig. 4.1a**). Localization signals included three commonly used

MTSs—COX8 (29 amino acid-long peptide derived from human cytochrome c oxidase subunit VIII)[34], Su9 (69 amino acid-long peptide derived from *Neurospora crassa* ATPase subunit 9)[35], and ATG4D (42 amino acid-long peptide derived from a Atg4 cysteine protease)[36]—that were previously used to probe the import of CRISPR nucleases into mammalian mitochondria[15]. The use of no localization signal or a nuclear localization signal (NLS) served as negative controls. Nineteen POIs were chosen across five different protein classes that have been used as components of precise gene editing technologies: Class I CRISPR systems (Cas3/CASCADE)[37], Class II CRISPR/Cas nucleases, RecTs, single-stranded binding proteins (SSBs), and retron reverse transcriptases (RTs). Both Class I and II CRISPR systems can cut DNA at programmable sites to induce editing using double-stranded break (DSB) repair pathways[38]. In contrast, RecTs and SSBs have been used to integrate donor DNA into bacterial and yeast genomes through recombineering[39–42]. Finally, retron RTs allow *in vivo* production of DNA donor editing templates in both prokaryotes and eukaryotes that mediate precise editing using either DSB repair pathways or recombineering[43–46]. As a positive control, we used the construct mito-APEX2, a protein that contains an MTS derived from the mitochondrially imported COX4 fused to APEX2, which has been shown to localize to the mammalian mitochondrial matrix using immunocytochemistry and proteomic mapping that found mito-APEX2 in close proximity to mitochondrial matrix proteins[47].

To engineer mammalian cell lines expressing a given protein cassette, each construct was cloned into a PiggyBac vector under the control of a doxycycline-inducible promoter adjacent to a constitutive puromycin resistance gene. These cassettes were randomly integrated into the genome of HEK293T cells using the PiggyBac transposase system and selected with puromycin (**Fig. 4.1b**). Biological replicates of a given construct were defined as either individual clones derived from a single bulk transposase integration or multiple parallel transposase integrations.

We screened the localization of each cassette by seeding cells into 96-well plates, expressed each protein for 24 hours under an inducible promoter, performed immunocytochemistry, and imaged each well using a high-throughput confocal microscope (ImageXpress Micro Confocal High-Content Imaging System). Specifically, each cell line was imaged for nuclei using Hoescht, POI using an antibody against HA, and mitochondria using an antibody against the mitochondrial marker TIM23 (**Fig 4.1c**). Using this high-throughput method, we found that localization of some cassettes varied qualitatively depending on localization signal. For instance, while the mito-APEX2 and Su9-LbCas12a showed punctate expression that colocalizes with mitochondria, NLS-LbCas12a showed clear nuclear colocalization and LbCas12a with no localization signal showed a diffuse, cytoplasmic phenotype. While these particular lines illustrate the expected localization based on signal, the localization of many other cassettes, such as of ATG4D-LbCas12a, were less predictable or more ambiguous. Thus, we next developed an analytical pipeline to quantify localization within mitochondria or nuclei.

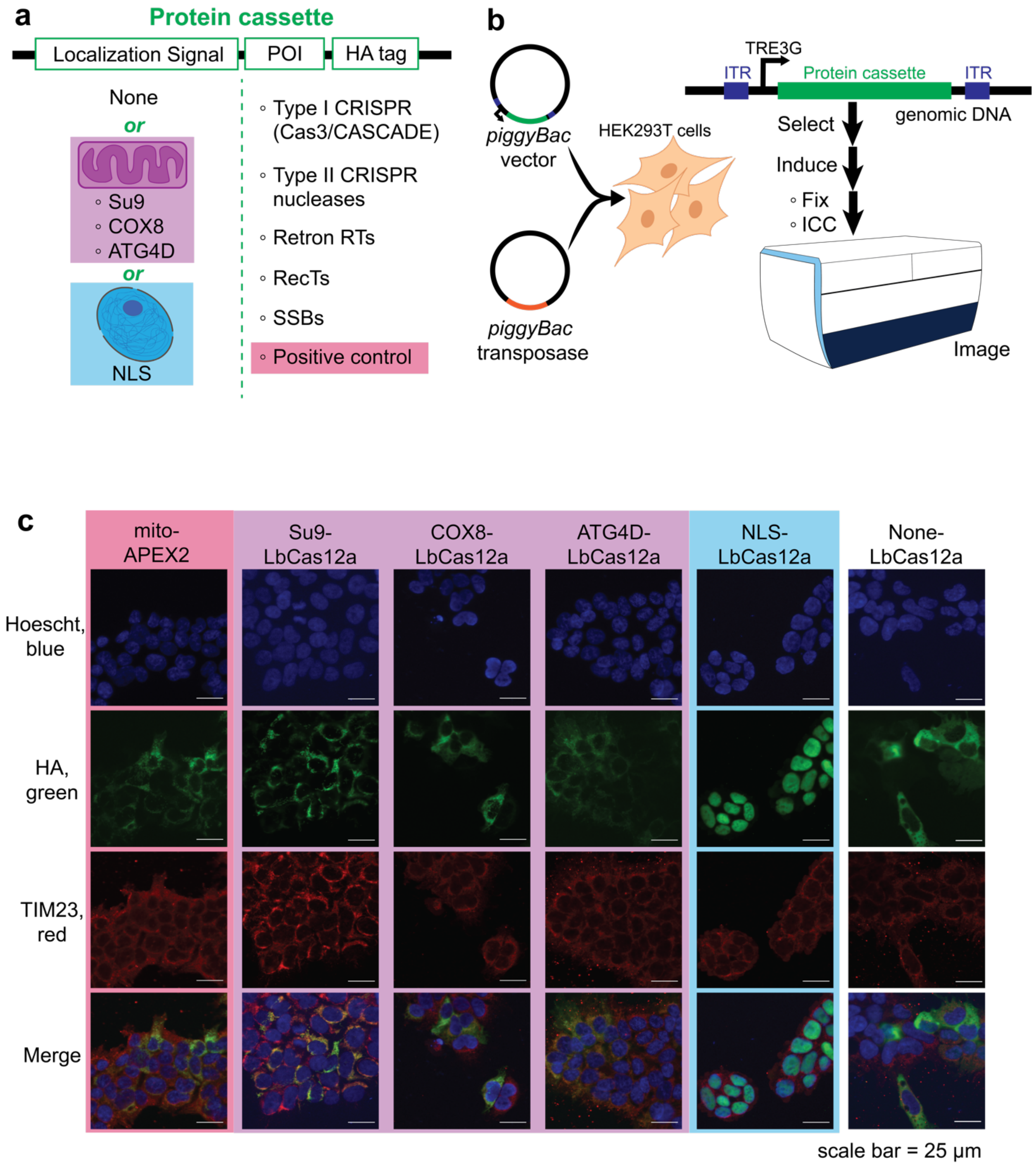


Figure 4.1 High-throughput localization workflow. (a) 66 genetic cassettes containing a combination of localization signal and POI, followed by a C-terminus HA tag, were synthesized. Localization signals include three N-terminus MTSs (Su9, COX8, and ATG4D), a nuclear localization signal (NLS), or no localization tag (None). POIs were chosen from five different protein families—Class I CRISPR/Cas proteins, Class II CRISPR nucleases, retron RTs, RecTs, and SSBs. A positive control of mito-APEX2 was also included. **(b)** Each cassette was randomly integrated into the genome of HEK293T cells using a piggyBac transposase system. Following co-

transfection of the cassette in a piggyBac vector and a plasmid constitutively expressing piggyBac transposase, cells were selected for at least one week using the antibiotic puromycin. To image each cell line, expression of a given cassette was induced for 24 hours using doxycycline before being fixed and stained prior to imaging using a high-throughput confocal microscope (ImageXpress Micro Confocal High-Content Imaging System). **(c)** Engineered cell lines were stained with Hoescht (blue), an antibody against HA (green), and an antibody against the mitochondrial marker TIM23 (red). Shown are representative images from the positive control (red background) and LbCas12a fused to one of three MTSs (purple), NLS (blue), or no localization tag (white).

4.2.2 Automated, quantitative, and open-source analysis pipeline

To better compare localization differences exhibited by MTS-POI combinations, we developed an unbiased Python-based analysis pipeline to quantify the mitochondrial and nuclear import between our dozens of cassettes. Crucially, we found that expression and localization were variable between individual cells of a given condition so our analysis pipeline is built to quantify colocalization at the level of single cells. Images corresponding to the nuclei and mitochondria for each biological replicate cell line from each condition were fed into a Cellpose-based machine learning model[48,49] to label individual cells (**Fig. 4.2a,b**). Next, cells were filtered using Otsu thresholding to remove any cells with no detectable protein expression (**Fig. 4.2c**). Specifically, Otsu thresholding was applied on each image to determine the pixel intensity threshold separating POI signal from background fluorescence. This analysis also revealed that some images had such low fluorescence that signal was effectively indistinguishable from noise. To ensure these specific images did not bias the final colocalization scores, any images in which Otsu thresholding did not separate signal from noise in each cell, as defined by the majority of filtered cells in an image failing to show a non-Gaussian intensity distribution typical of true fluorescent signal, were discarded (**Supplemental Figure S1**). In some cases, only a few images for a cell line were eliminated, although—in cases where a clonal or transfected line suffered from minimal cassette expression—the entire biological replicate was removed from analysis.

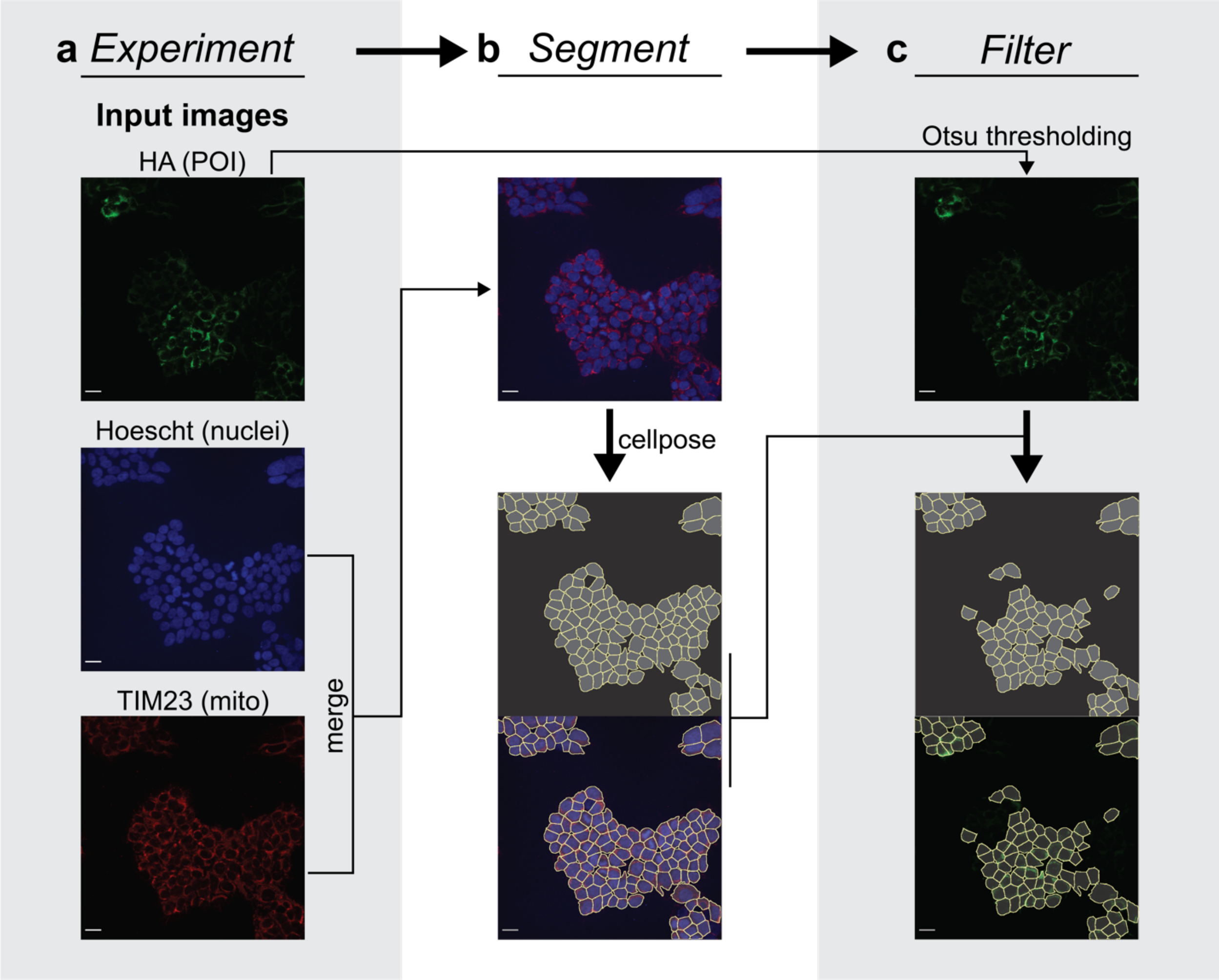


Figure 4.2 Machine learning algorithm enables automated labeling of cells containing cassette protein. Scale bar = 25 µm. **(a)** Unprocessed images from each fluorescent channel are fed into a Python-based analysis pipeline to segment individual cells. Shown are representative images from the cell line Su9-LbCas12a (top, HA; middle; Hoescht, bottom; TIM23). **(b)** The Hoescht and TIM23 channels are merged (top) prior to being fed into a custom, retrained neural network (arrow), resulting in automated labeling of each cell found within the image. Images below arrow; top image shows Cellpose-generated mask of segmented cells, where each yellow line indicates a cell boundary. Bottom image shows mask overlaid atop merged Hoescht/TIM23 channels. **(c)** Cellpose-segmented cells are filtered to keep only cells which express the cassette protein. Otsu thresholding is applied to the HA channel (top) to determine a threshold separating true fluorescent signal from background noise. Segmented cells containing at least 50 pixels with signal are kept to perform further colocalization measurements. Images below arrow; top image shows segmented cells following filtering. Bottom image shows filtered mask overlaid atop HA channel.

After selecting a population of filtered cells to further analyze for a given protein cassette, colocalization between the cassette protein and either mitochondria (**Fig. 4.3a,b**) or nuclei (**Fig. 4.3d,e**) was measured on a per cell basis using Pearson's correlation coefficient (PCC)[50]. Using this method, PCC scores vary between -1 (anti-correlated) to +1 (highly correlated). High colocalization scores indicate that a protein is collocated with a given organelle while low colocalization scores suggest little to no specific colocalization between a POI and a given organelle occurred.

For our analysis, we considered individual cells that survived quality filters from a single transfection or clone as technical replicates and summarized the overall colocalization score for a single biological replicate of each protein cassette by taking the median of all the individual cell colocalization scores for mitochondria (**Fig. 4.3c**) or nuclei (**Fig. 4.3f**). We replicated our experiments using at least three different transfections or five clonal lines as biological replicates.

We generally found low variability within our biological replicates, suggesting that protein import is a fairly reliable phenomenon. The positive control mito-APEX2 obtained an average score of 0.63 +/- 0.03 (mean +/- std. dev), a high PCC value that strongly implies mito-APEX2 is imported into the mitochondrial matrix. Similar to our previous qualitative assessments (**Fig 4.1c**), we found that colocalization scores, even across a single POI, vary depending on localization signals. The colocalization scores of LbCas12a fused to the Su9 or COX8 MTSs were not statistically different from mito-APEX2, suggesting mitochondrial import had occurred. In comparison, when LbCas12a was instead fused to ATG4D, no localization signal, or NLS, colocalization scores dropped significantly, indicating less mitochondrial import occurred (**Fig. 4.3g**). Moreover, when comparing nuclear colocalization scores (**Fig. 4.3h**), all cell lines except NLS-LbCas12a showed a consistent, low nuclear colocalization score, suggesting little to no nuclear import. As expected, only the cell line fused to a NLS had a high colocalization score indicating high nuclear import. These findings suggest that our workflow is able to compare import efficiencies across different combinations of MTS and POI for multiple organelles.

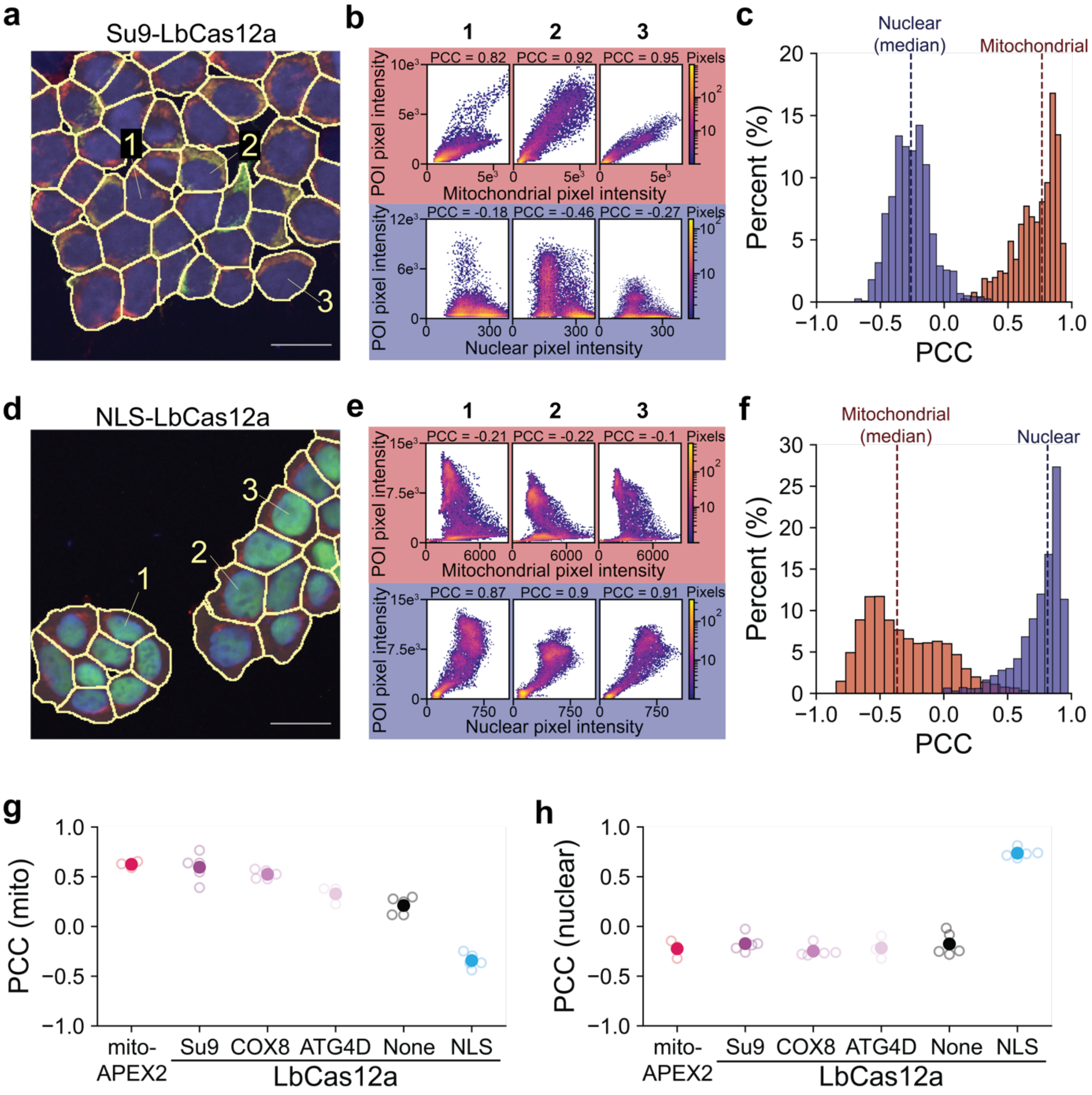


Figure 4.3 Computational workflow quantifies the colocalization of cassette proteins with mitochondria or nuclei. (a) Representative image of a cell line (Su9-LbCas12a) with clear mitochondrial expression of its cassette protein, with mask in yellow overlaid on top. Numbers refer to three representative cells for which data is shown in (b). Scale bar = 25 µm. **(b)** Heatmaps depicting the relationship between Su9-LbCas12a pixel intensity and organellar pixel intensity (mitochondria on top in red; nuclei on bottom in blue) for each pixel within a representative cell

from (a) (left; cell #1, middle; cell #2, right; cell #3). Color depicts the number of pixels. The strength of the linear relationship between pixel intensities, or colocalization, within each cell is calculated using PCC, and the result depicted on top its respective heatmap. **(c)** Histogram depicting the all the colocalization scores for all the cells for one clonal line expressing Su9-LbCas12a. Mitochondrial PCC scores are shown in red, while nuclear PCC scores are shown in blue. Dotted lines depict the median colocalization score for mitochondria (red) and nuclei (blue). **(d)** Representative image of a cell line (NLS-LbCas12a) with clear nuclear expression of its cassette protein, with mask in yellow overlaid on top. Numbers refer to three representative cells. Scale bar = 25 μm. **(e)** Heatmaps depicting the relationship between NKS-LbCas12a pixel intensity and organellar pixel intensity (mitochondria on top in red; nuclei on bottom in blue) for each pixel within a representative cell from (d) (left; cell #1, middle; cell #2, right; cell #3). Color depicts the number of pixels. The strength of the linear relationship between pixel intensities, or colocalization, within each cell is calculated using PCC, whose result is on top its respective heatmap. **(f)** Histogram depicting the all the colocalization scores for all the cells for one clonal line expressing NLS-LbCas12a. Mitochondrial PCC scores are shown in red, while nuclear PCC scores are shown in blue. Dotted lines depict the median colocalization score for mitochondria (red) and nuclei (blue). **(g)** Mitochondrial colocalization of positive control mito-APEX2 (red) and LbCas12a fused to different localization tags, as measured using the described experimental and analytic workflow. There is a significant effect of localization signal (one-way ANOVA, P<0.0001), where ATG4D (P=0.0021), no signal (P<0.0001), and NLS (P<0.0001) are all significantly different from mito-APEX2, but Su9 (P=0.9831) and COX8 (P=0.376) are not (Dunnett's corrected). Open circles are biological replicates; closed circles are average of all biological replicates. **(h)** Nuclear colocalization of mito-APEX2 (red) and LbCas12a fused to different localization tags measured using the described experimental and analytical workflow. There is a significant effect of localization signal (one-way ANOVA, P<0.0001), where NLS (P<0.0001) is significantly different from mito-APEX2, but Su9 (P=0.8587), COX8 (P=0.9930), ATG4D (P=0.9998), and no signal (P=0.0.8825) are not (Dunnett's corrected). Open circles are biological replicates; closed circles are average of all biological replicates.

4.2.3 MTS selection strongly influences mitochondrial import

After validating the analytical pipeline, we used this workflow to quantify the mitochondrial and nuclear import of all 66 different protein cassettes (**Fig. 4.4a; Supplemental Figure S2a,b**). Interestingly, mitochondrial colocalization scores did not cluster bimodally into high and low scores. Instead, scores were distributed continuously, suggesting that different cassettes have varying capabilities to drive POIs to the mitochondria.

Previous studies have found that both the specific MTS and hydrophobicity of individual POIs can influence the probability that a MTS-POI will localize in the mitochondria[31,19,33]. To determine if localization signal influences mitochondrial import across the four protein families tested, we investigated how the mitochondrial import rank of cassettes clustered based on localization signal (**Fig. 4.4b**). We fit a linear mixed effects model with a fixed effect of localization signal and a random intercept within each protein family, to specifically test the statistical significance of each localization signal while accounting for the variability inherent within each protein of interest. As expected, we found that cassettes clustered based on specific localization signals; while fusion to an MTS resulted in higher rankings across the board, having a NLS led to clear clustering at the bottom of the rankings. However, there was a clear ranking priority to how well each MTS performed compared to each other, with Su9 and COX8 driving high mitochondrial import, while ATG4D performed significantly worse.

Given ATG4D's uneven performance across different POIs, we next further analyzed how colocalization scores varied based on localization scores within different protein families (**Fig. 4.4c-g**). Interestingly, all three MTSs were able to drive high mitochondrial import of Class I CRISPR-related proteins (**Fig. 4.4c**), whereas both COX8 and ATG4D appeared less capable than Su9 of importing the larger Class II CRISPR nucleases (**Fig. 4.4d**). Of the two retron RTs tested, only Su9 led to mitochondrial localization of retron-Eco1 RT, whereas COX8 and ATG4D were able to localize retron Eco2 RT (**Fig. 4.4e**). Finally, while both COX8 and Su9 were able to import all RecTs and SSBs tested into the mitochondria (**Fig. 4.4f,g**), ATG4D instead appeared to misdirect these proteins to a different location, based on an unusual punctate pattern which did not colocalize with mitochondria (**Supplemental Figure S2c**).

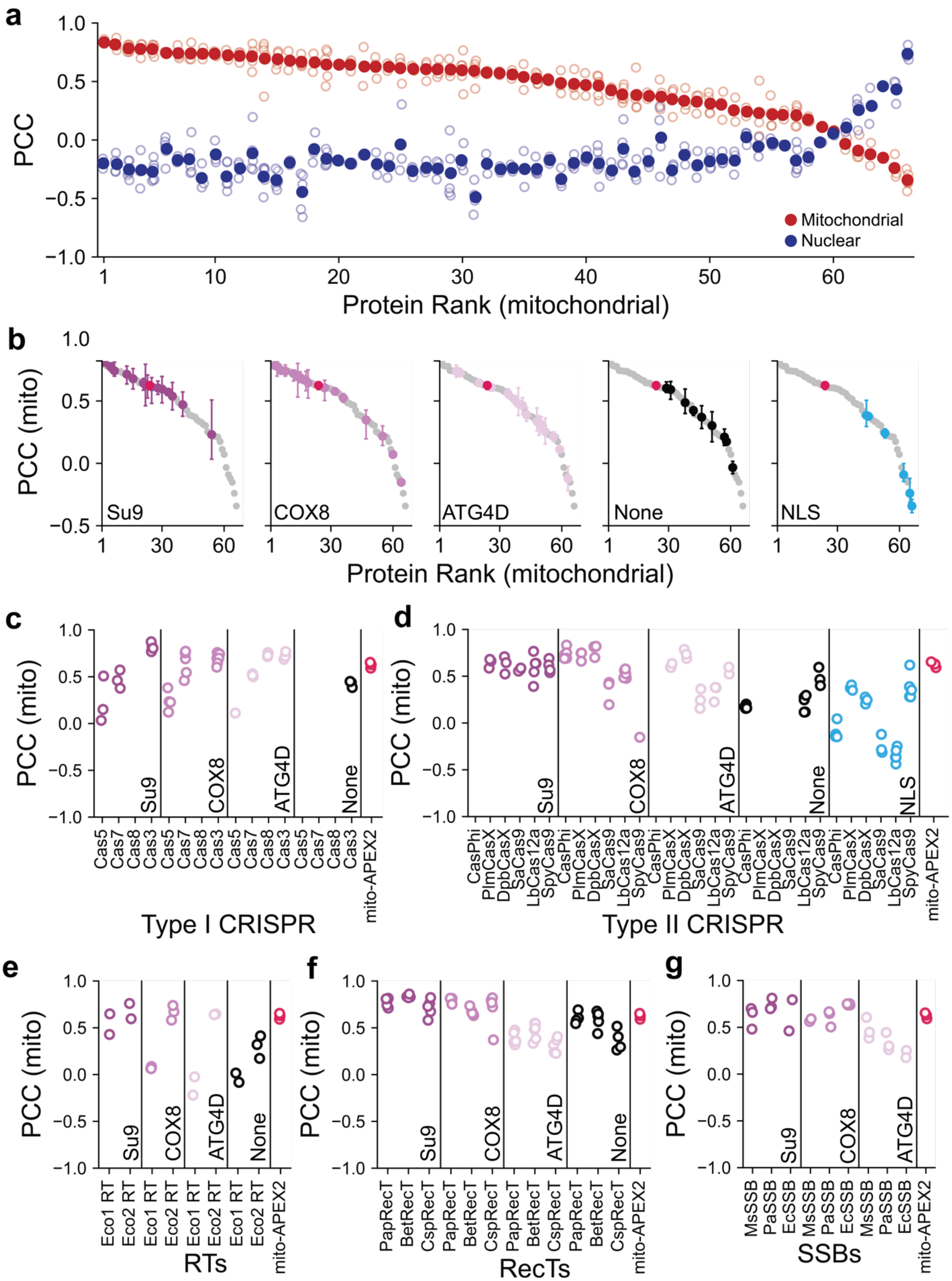

a
PCC
1.0
0.5
0.0
−0.5
−1.0
● Mitochondrial
● Nuclear
1
10
20
30
40
50
60
Protein Rank (mitochondrial)
b
PCC (mito)
1.0
0.5
0.0
−0.5
Su9
COX8
ATG4D
None
NLS
1
30
60
Protein Rank (mitochondrial)
c
PCC (mito)
Su9
COX8
ATG4D
None
Cas5
Cas7
Cas8
Cas3
mito-APEX2
Type I CRISPR
d
PCC (mito)
Su9
COX8
ATG4D
None
NLS
CasPhi
PlmCasX
DpbCasX
SaCas9
LbCas12a
SpyCas9
mito-APEX2
Type II CRISPR
e
PCC (mito)
Su9
COX8
ATG4D
None
Eco1 RT
Eco2 RT
mito-APEX2
RTs
f
PCC (mito)
Su9
COX8
ATG4D
None
PapRecT
BetRecT
CspRecT
mito-APEX2
RecTs
g
PCC (mito)
Su9
COX8
ATG4D
MsSSB
PaSSB
EcSSB
mito-APEX2
SSBs

Figure 4.4 MTS selection strongly influences mitochondrial import. (a) Mitochondrial (red) and nuclear (blue) colocalization scores of 66 different protein cassettes. Proteins were ranked based on average mitochondrial colocalization score, from highest mitochondrial PCC to lowest. Open circles are biological replicates; closed circles are average of all biological replicates. **(b)** Clustering of cassettes based on specific localization signal. The entire continuum of all mitochondrial colocalization scores from (a) is shown in each subplot in gray, while the positive control mito-APEX2 is shown in each subplot in red. The clustering of cassettes with a specific localization signal is shown in another color on top (from left to right subplots: Su9, COX8, ATG4D, none, and NLS). There is a significant effect of localization signal (linear mixed effects model, $P<0.0001$), where ATG4D is significantly different from both Su9 ($P<0.0001$) and COX8 ($P<0.0001$) but not None (0.092559) during follow-up (Kenward-Roger corrected). Meanwhile, there is no significant difference between Su9 and COX8 ($P=0.88139$). Closed circles indicate average of all biological replicates; error bars indicate standard deviation. **(c)-(g)** Mitochondrial colocalization scores broken down by localization signal within a given protein family (**c**; Class I CRISPR Cas3/CASCADE, **d**; Class II CRISPR nucleases, **e**; retron RTs, **f**; RecTs, **g**; SSBs). Specific proteins are listed in order of amino acid length, from shortest on the left to longest on the right. As a reference for mitochondrial PCC scores suggestive of mitochondrial import, scores for positive control mito-APEX2 are shown to the right of each figure (red). Open circles are individual biological replicates.

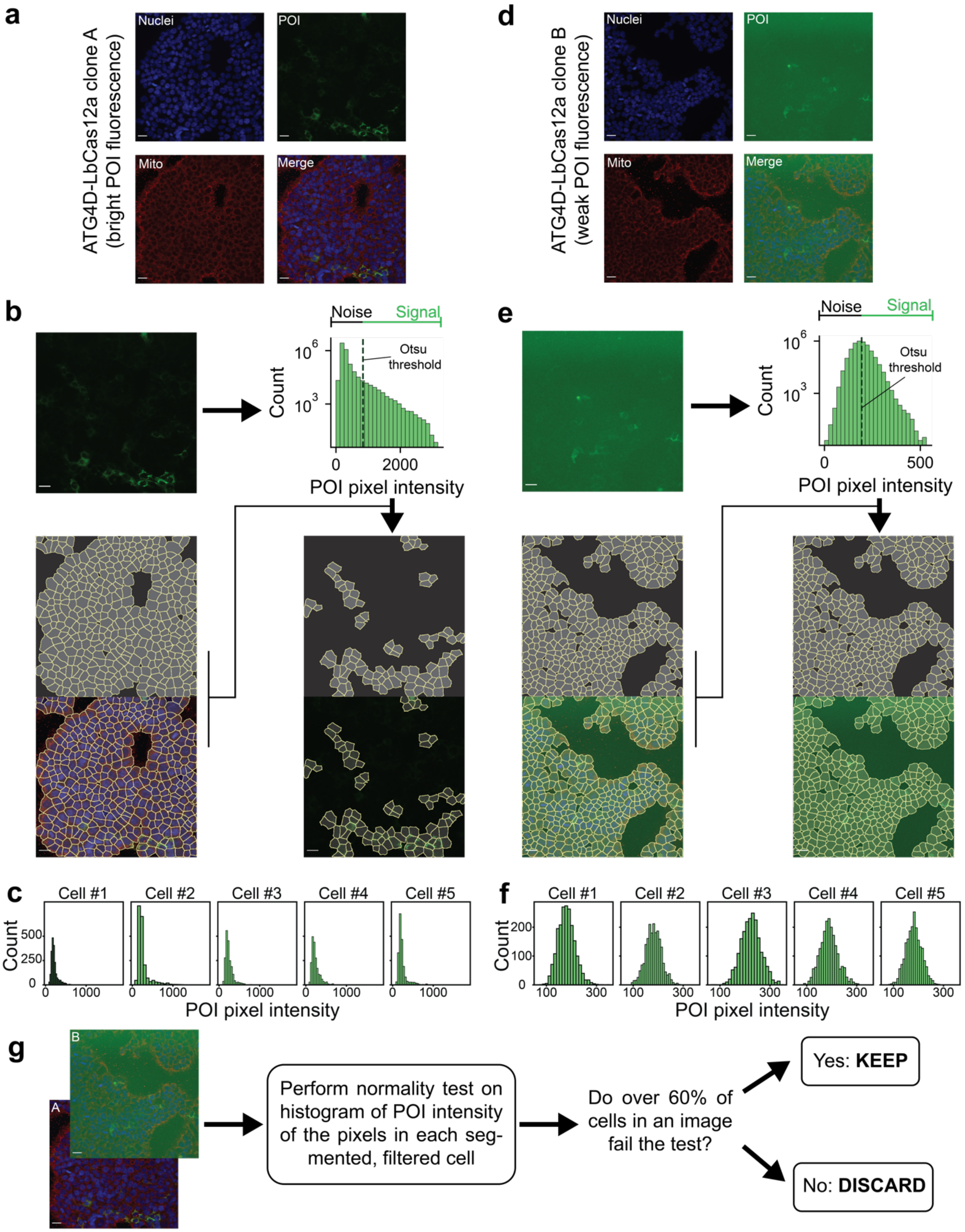
4.3 Discussion

Importing non-canonical proteins into mammalian mitochondria is a critical step to overcome particular challenges in metabolic engineering and for developing therapeutics to treat genetic diseases of the mtDNA. Here, we design a high-throughput imaging-based workflow to quickly screen the subcellular localization of a tagged protein. This method enabled us to determine the best MTS-POI combination across three commonly used MTSs and five different protein classes that are components of gene editing technologies. The results from these screens should help drive more robust mitochondrial gene editing by enabling researchers to test mitochondrial gene editing using other nucleases or proteins beyond SpyCas9, which has been reported to cause mitochondrial dysfunction when imported to mitochondria[15], or testing MTSs that yield more specific mitochondria import to prevent off-target effects in nuclear DNA[29].

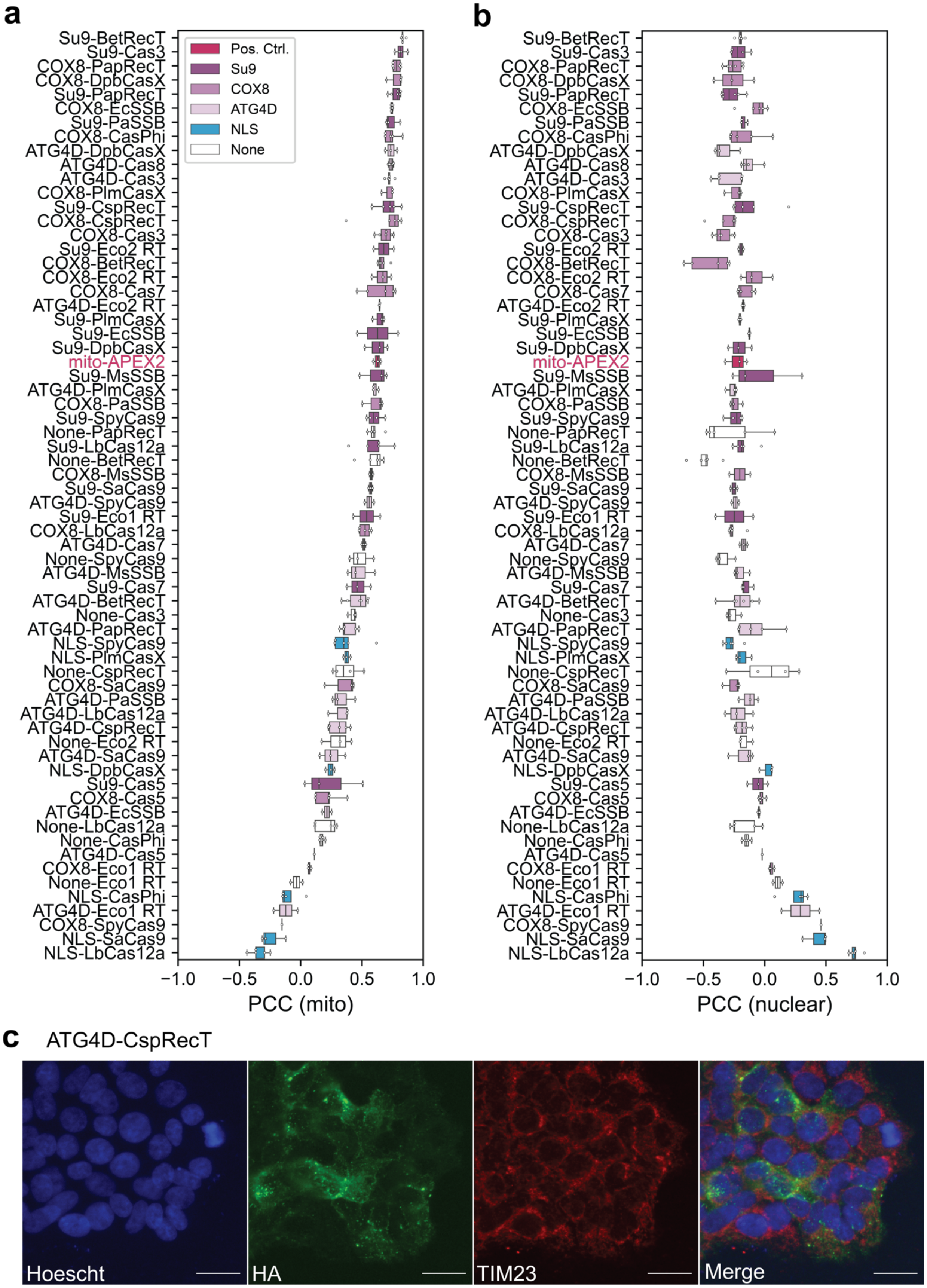
This work also reveals broader trends about which MTS to use across multiple protein types. By testing three common MTSs on a diverse set of POIs, we find that Su9 and COX8 consistently performed the best while ATG4D performed the most unevenly and misdirected several POIs to an alternate location. In addition, unlike COX8 and ATG4D, Su9 was able to import specific proteins, such as large Class II nucleases and Eco1 RT, into the mitochondria.

As a tool for the field, our Python-based workflow implemented in an annotated Jupyter notebook can be reused for future experiments, including more generally to other screens using different proteins, organelles of interest, or cell lines. Although we used a high-throughput confocal microscope in our own workflow, other confocal microscopes could easily be used, depending on the necessary throughput or number of samples. Computationally, the analytic pipeline uses Python, an open-source programming language, and relies on pixel-based colocalization analyses and quality check steps to eliminate cells not expressing a protein of interest that can be universally applied regardless of cell line or differing expression levels or phenotypes[50]. The only experiment-specific alteration to the pipeline would be to apply a different neural network to segment individual cells, depending on the cell line and seeding density used. However, Cellpose already offers a collection of ready-to-use neural networks or—if further refinement is necessary—an intuitive GUI that enables users to retrain and create their own neural network in fewer than 30 minutes[49]. Therefore, others should be able to easily apply this analytical framework to quantify colocalization within their own fluorescent images.

4.4 Supplementary Figures



Supplemental Figure S1, related to Figure 4.2 in the main text. Normality test implemented in Python to remove cell lines with low signal-to-noise ratio of cassette protein, as part of a computational workflow to quantify the import of cassette proteins into mitochondria or nuclei. All scale bars = 25 µm. **(a)** Representative image of a clonal line of ATG4D-LbCas12a with strong expression of its cassette protein across three fluorescent channels (Hoescht, TIM23, and HA) and a merge of all channels. **(b)** Segmenting and filtering using Otsu threshold for the representative image in (a). Otsu thresholding is applied to determine pixel intensity threshold that separates "noise" from "signal" or true fluorescence. This threshold is used to remove all individual cell masks that do not have >50 pixels of "signal" to create a smaller sub-selection of filtered cells to analyze. **(c)** To check that the filtering algorithm appropriately removed cells without signal, the shape of the histogram of POI intensity pixels per filtered cell is checked for normality. Here, give representative cells from the filtered cells chosen in (b) are shown. Note all histograms show a skewed, non-normal shape that is indicative of true signal. **(d)** Representative image of clonal line of ATG4D-LbCas12a with weak expression of its cassette protein across three fluorescent channels (Hoescht, TIM23, and HA) and a merge of all channels. **(e)** The same segmenting and filtering process and described in (b) is shown for the representative image from (d). Note that the chosen Otsu threshold does not eliminate any cells from the mask. **(f)** The histograms of POI pixel intensity from five representative cells from the filtered cells from (e) are shown. Note all histograms show a normal distribution, indicative of noise rather than true signal. **(g)** Schematic showing the implementation of normality test to remove cell lines with low signal-to-noise ratio of cassette protein. A normality test on the histogram of POI intensity in each filtered cell is performed per image. If over 60% of the cells in an image are non-normal, indicative of true signal, the image is kept while. If over 60% of the cells are normal, then the image is instead discarded and not included in future analyses.


a
Pos. Ctrl.
Su9
COX8
ATG4D
NLS
None
Su9-BetRecT
Su9-Cas3
COX8-PapRecT
COX8-DpbCasX
Su9-PapRecT
COX8-EcSSB
Su9-PaSSB
COX8-CasPhi
ATG4D-DpbCasX
ATG4D-Cas8
ATG4D-Cas3
COX8-PlmCasX
Su9-CspRecT
COX8-CspRecT
COX8-Cas3
Su9-Eco2 RT
COX8-BetRecT
COX8-Eco2 RT
COX8-Cas7
ATG4D-Eco2 RT
Su9-PlmCasX
Su9-EcSSB
Su9-DpbCasX
mito-APEX2
Su9-MsSSB
ATG4D-PlmCasX
COX8-PaSSB
Su9-SpyCas9
None-PapRecT
Su9-LbCas12a
None-BetRecT
COX8-MsSSB
Su9-SaCas9
ATG4D-SpyCas9
Su9-Eco1 RT
COX8-LbCas12a
ATG4D-Cas7
None-SpyCas9
ATG4D-MsSSB
Su9-Cas7
ATG4D-BetRecT
None-Cas3
ATG4D-PapRecT
NLS-SpyCas9
NLS-PlmCasX
None-CspRecT
COX8-SaCas9
ATG4D-PaSSB
ATG4D-LbCas12a
ATG4D-CspRecT
None-Eco2 RT
ATG4D-SaCas9
NLS-DpbCasX
Su9-Cas5
COX8-Cas5
ATG4D-EcSSB
None-LbCas12a
None-CasPhi
ATG4D-Cas5
COX8-Eco1 RT
None-Eco1 RT
NLS-CasPhi
ATG4D-Eco1 RT
COX8-SpyCas9
NLS-SaCas9
NLS-LbCas12a
−1.0
−0.5
0.0
0.5
1.0
PCC (mito)
b
Su9-BetRecT
Su9-Cas3
COX8-PapRecT
COX8-DpbCasX
Su9-PapRecT
COX8-EcSSB
Su9-PaSSB
COX8-CasPhi
ATG4D-DpbCasX
ATG4D-Cas8
ATG4D-Cas3
COX8-PlmCasX
Su9-CspRecT
COX8-CspRecT
COX8-Cas3
Su9-Eco2 RT
COX8-BetRecT
COX8-Eco2 RT
COX8-Cas7
ATG4D-Eco2 RT
Su9-PlmCasX
Su9-EcSSB
Su9-DpbCasX
mito-APEX2
Su9-MsSSB
ATG4D-PlmCasX
COX8-PaSSB
Su9-SpyCas9
None-PapRecT
Su9-LbCas12a
None-BetRecT
COX8-MsSSB
Su9-SaCas9
ATG4D-SpyCas9
Su9-Eco1 RT
COX8-LbCas12a
ATG4D-Cas7
None-SpyCas9
ATG4D-MsSSB
Su9-Cas7
ATG4D-BetRecT
None-Cas3
ATG4D-PapRecT
NLS-SpyCas9
NLS-PlmCasX
None-CspRecT
COX8-SaCas9
ATG4D-PaSSB
ATG4D-LbCas12a
ATG4D-CspRecT
None-Eco2 RT
ATG4D-SaCas9
NLS-DpbCasX
Su9-Cas5
COX8-Cas5
ATG4D-EcSSB
None-LbCas12a
None-CasPhi
ATG4D-Cas5
COX8-Eco1 RT
None-Eco1 RT
NLS-CasPhi
ATG4D-Eco1 RT
COX8-SpyCas9
NLS-SaCas9
NLS-LbCas12a
−1.0
−0.5
0.0
0.5
1.0
PCC (nuclear)
c
ATG4D-CspRecT
Hoescht
HA
TIM23
Merge

Supplemental Figure S2, related to Figure 4.4 in the main text. Summary of the import of protein cassettes using our quantitative, high-throughput pipeline. (a) Mitochondrial colocalization scores of 66 different protein cassettes. Proteins are arranged from top to bottom based on average mitochondrial colocalization score, from highest mitochondrial PCC to lowest. Open circles are technical replicates. **(b)** Nuclear colocalization scores of 66 different protein cassettes. Proteins are arranged from top to bottom based on average mitochondrial colocalization score, from highest mitochondrial PCC to lowest. Open circles are technical replicates. **(c)** Representative image of a clonal line of ATG4D-CspRecT across three fluorescent channels (Hoescht, HA, and TIM23) and a merge of all channels. Note the punctate expression of HA that does not align with mitochondria (TIM23).

4.5 Methods

4.5.1 Constructs and strains

Protein cassettes were constructed by amplifying localization signals and POI nucleotide sequences using PCR from synthesized gBlocks (IDT) or existing plasmids. Complete protein cassettes were cloned into a PiggyBac integrating plasmid for doxycycline-inducible human protein expression (TetOn-3G promoter) using Gibson assembly. Alternatively, some cassettes were synthesized into the same custom PiggyBac integrating plasmid by Twist Bioscience.

Stable mammalian cells for imaging were generated using the standard Lipofectamine 3000 transfection protocol (Invitrogen) and a PiggyBac transposase system. T12.5 flasks with 50-70% confluent HEK293T cells were transfected using 1.6 µg POI cassette expression plasmid and 0.8 µg PiggyBac transposase plasmid (pCMV-hyPBase). Stable cell lines were selected using puromycin for at least one week.

Clonal lines were generated by growing individual cells into separate cell populations. Specifically, stable cell lines were serially diluted to a final concentration of 2.5 cells per mL media then seeded into a 96-well plate using 100 µL/well. Wells that received a single cell had media refreshed weekly until a clonal line proliferated to ~40% confluency, at which point a clonal line was passaged to a larger flask for further experiments.

4.5.2 Immunocytochemistry

96-well glass bottom plates with #1 cover glass (Cellvis, catalog # P96-1-N) were coated with a mixture of 50% poly-D-lysine (ThermoFisher Scientific, catalog #A3890401) and DPBS (ThermoFisher Scientific, catalog #14040133) for 30 minutes at room temperature. Wells were washed three times with distilled water and left out to dry for at least 2 hours prior to seeding.

Cells were seeded at a density of 10,000 cells per well. The following day, doxycycline was added at a final concentration of 1 µg/mL to induce expression of the protein cassette. At 24 hours post-induction, cell nuclei were stained using a final concentration of 10 µM Hoescht for at least 5 minutes prior to fixation.

For fixation, media was aspirated from each well and replaced with a solution of 4% paraformaldehyde (PFA) created fresh by fixing a 1 mL 16% (w/v) PFA ampule (ThermoFisher Scientific, catalog #28906) with 3 mL PBS. Cells were fixed for 30 minutes at room temperature prior to three 5-minute washes with PBS. Following fixation, cells were permeabilized and blocked for an hour at room temperature using blocking buffer made fresh with the following ingredients: PBS containing 10% donkey serum (Sigma-Aldrich, catalog #D9663), 10% Triton X-100 (Sigma-Aldrich, catalog #X100), and 100 mg BSA (Sigma-Aldrich, catalog #A9418) per 10 mL solution. Next, cells were incubated overnight at 4°C in blocking buffer with the antibodies anti-HA tag conjugated to DyLight 550 (ThermoFisher Scientific, catalog #26183-D550) and anti-TIM23 (Abcam, catalog #ab230253) each added at a 1:100 dilution. After performing three more 5-minute washes, cells were incubated with a secondary antibody goat anti-rabbit conjugated to DyLight 650 (ThermoFisher Scientific, catalog #84546) at a 1:500 dilution in blocking buffer for 3 hours. Following secondary antibody incubation, three more 5-minute washes were performed prior to the addition of 30 uL antifade mountant (ThermoFisher Scientific, catalog #S36967) per well.

Plates were wrapped in aluminum foil to avoid light and either stored temporarily at 4°C or at -20°C for longer-term storage prior to imaging.

4.5.3 Imaging

Stained cells were imaged using an ImageXpress Micro Confocal High-Content Imaging System (Molecular Devices) using a 40X water immersion objective by taking a 7-layer Z-stack, with each layer spaced 0.3 µm apart, at four different sites per well.

4.5.4 Colocalization Image Analysis Pipeline

A colocalization image analysis pipeline was made using jupyter-notebook in Python 3[51,52], and uses the following packages: numpy, pandas, scipy, skimage, tdqm, and tifffile. Additionally, the pipeline requires the Cellpose code library[48,49] along with these additional packages: numba, opencv, and pytorch. Using the Cellpose GUI also requires PyQt and pyqtgraph[53].

TIFF files consisting of merged nuclear and mitochondrial channels were created using a custom function and fed into a neural network retrained according to the instructions for the Cellpose GUI[49]. Briefly, the "CP" model from the Cellpose model zoo was initially used to segment all images. Afterwards, about five images with poor initial segmentation were chosen for manual annotation. The CP model was then retrained using the corrected labels and the new model was re-run on all images.

To remove segmented cells that did not contain expression of the cassette protein, Otsu thresholding was performed on the HA channel to determine a pixel intensity threshold separating signal from background for each image. Only segmented cells containing at least 50 pixels of cassette protein signal, referred to as filtered cells, were kept for further analysis.

Two additional functions to ensure quality-check steps were also implemented. First, any image containing fewer than six filtered cells was automatically removed from further analysis. Second, since the overall expression of a cassette protein can vary between different cell lines, a function was written to ensure that the filtering step effectively distinguished between cells that did or did not express a cassette protein. Individual cells containing noise, rather than signal, exhibit a Gaussian distribution of protein cassette pixel intensities. In contrast, cells with signal tend to exhibit non-normal or skewed pixel intensity distributions. Thus, for every image, a “non-Gaussian” test was performed on each filtered cell by testing for normality. If over 60% of filtered cells failed the “non-Gaussian” test, then this result suggests that the majority of filtered cells within the image do not contain true expression of the cassette protein, thus that specific image would be removed from further analysis.

Afterwards, a custom function was built to calculate PCC between the HA channel pixel intensities and either the mitochondrial or nuclear channel pixel intensities for every filtered cell related to a given biological replicate. Due to the skew present in most PCC distribution, these results were summarized by taking the median of all the filtered cells for a given biological replicate.

4.5.5 Statistics

ANOVA and post-hoc analyses were performed using GraphPad Prism v9.4.1 For linear mixed effects modeling, we used *R* v4.2.3 using the *lme4* and *lmerTest* packages (post-hoc analyses used the Kenward-Roger degrees of freedom correction method implemented in the package *pbkrtest*).

4.6 References

1. Agrimi, G. Role of Mitochondrial Carriers in Metabolic Engineering. *J. Bioprocess. Biotech.* **04**, (2014).
2. Huttanus, H. M. & Feng, X. Compartmentalized metabolic engineering for biochemical and biofuel production. *Biotechnol. J.* **12**, 1700052 (2017).
3. Verechshagina, N. A., Konstantinov, Yu. M., Kamenski, P. A. & Mazunin, I. O. Import of Proteins and Nucleic Acids into Mitochondria. *Biochem. Mosc.* **83**, 643–661 (2018).
4. Di Donfrancesco, A. *et al.* Gene Therapy for Mitochondrial Diseases: Current Status and Future Perspective. *Pharmaceutics* **14**, 1287 (2022).
5. Farhi, M. *et al.* Harnessing yeast subcellular compartments for the production of plant terpenoids. *Metab. Eng.* **13**, 474–481 (2011).
6. Avalos, J. L., Fink, G. R. & Stephanopoulos, G. Compartmentalization of metabolic pathways in yeast mitochondria improves the production of branched-chain alcohols. *Nat. Biotechnol.* **31**, 335–341 (2013).
7. Taylor, R. W. & Turnbull, D. M. Mitochondrial DNA mutations in human disease. *Nat. Rev. Genet.* **6**, 389 (2005).
8. Ng, Y. S. & Turnbull, D. M. Mitochondrial disease: genetics and management. *J. Neurol.* **263**, 179–191 (2016).
9. Artika, I. M. Allotopic expression of mitochondrial genes: Basic strategy and progress. *Genes Dis.* **7**, 578–584 (2020).
10. Bacman, S. R., Williams, S. L., Pinto, M., Peralta, S. & Moraes, C. T. Specific elimination of mutant mitochondrial genomes in patient-derived cells by mitoTALENs. *Nat. Med.* **19**, 1111–1113 (2013).

11. Gammage, P. A., Rorbach, J., Vincent, A. I., Rebar, E. J. & Minczuk, M. Mitochondrially targeted ZFNs for selective degradation of pathogenic mitochondrial genomes bearing large-scale deletions or point mutations. *EMBO Mol. Med.* **6**, 458–466 (2014).

12. Reddy, P. *et al.* Selective Elimination of Mitochondrial Mutations in the Germline by Genome Editing. *Cell* **161**, 459–469 (2015).

13. Jo, A. *et al.* Efficient Mitochondrial Genome Editing by CRISPR/Cas9. *BioMed Res. Int.* **2015**, 305716 (2015).

14. Hashimoto, M. *et al.* MitoTALEN: A General Approach to Reduce Mutant mtDNA Loads and Restore Oxidative Phosphorylation Function in Mitochondrial Diseases. *Mol. Ther.* **23**, 1592–1599 (2015).

15. Antón, Z. *et al.* Mitochondrial import, health and mtDNA copy number variability seen when using type II and type V CRISPR effectors. *J. Cell Sci.* **133**, jcs248468 (2020).

16. Mok, B. Y. *et al.* A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* (2020) doi:10.1038/s41586-020-2477-4.

17. Hussain, S.-R. A., Yalvac, M. E., Khoo, B., Eckardt, S. & McLaughlin, K. J. Adapting CRISPR/Cas9 System for Targeting Mitochondrial Genome. *Front. Genet.* **12**, (2021).

18. Bi, R. *et al.* Direct evidence of CRISPR-Cas9-mediated mitochondrial genome editing. *The Innovation* **3**, (2022).

19. Chacinska, A., Koehler, C. M., Milenkovic, D., Lithgow, T. & Pfanner, N. Importing Mitochondrial Proteins: Machineries and Mechanisms. *Cell* **138**, 628–644 (2009).

20. Wiedemann, N. & Pfanner, N. Mitochondrial Machineries for Protein Import and Assembly. *Annu. Rev. Biochem.* **86**, 685–714 (2017).

21. Fukasawa, Y. *et al.* MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage Sites. *Mol. Cell. Proteomics MCP* **14**, 1113–1126 (2015).

22. Savojardo, C., Martelli, P. L., Fariselli, P. & Casadio, R. TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics* **31**, 3269–3275 (2015).

23. Almagro Armenteros, J. J. *et al.* Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* **2**, e201900429 (2019).

24. Bayne, A. N., Dong, J., Amiri, S., Farhan, S. M. K. & Trempe, J.-F. MTSviewer: a database to visualize mitochondrial targeting sequences, cleavage sites, and mutations on protein structures. 2021.11.25.470064 Preprint at https://doi.org/10.1101/2021.11.25.470064 (2022).

25. Van Steeg, H., Oudshoorn, P., Van Hell, B., Polman, J. E. & Grivell, L. A. Targeting efficiency of a mitochondrial pre-sequence is dependent on the passenger protein. *EMBO J.* **5**, 3643–3650 (1986).

26. Oca-Cossio, J., Kenyon, L., Hao, H. & Moraes, C. T. Limitations of Allotopic Expression of Mitochondrial Genes in Mammalian Cells. *Genetics* **165**, 707–720 (2003).

27. Perales-Clemente, E., Fernández-Silva, P., Acín-Pérez, R., Pérez-Martos, A. & Enríquez, J. A. Allotopic expression of mitochondrial-encoded genes in mammals: achieved goal, undemonstrated mechanism or impossible task? *Nucleic Acids Res.* **39**, 225–234 (2011).

28. Bader, G. *et al.* Assigning mitochondrial localization of dual localized proteins using a yeast Bi-Genomic Mitochondrial-Split-GFP. *eLife* **9**, e56649 (2020).

29. Wei, Y. *et al.* Mitochondrial base editor DdCBE causes substantial DNA off-target editing in nuclear genome of embryos. *Cell Discov.* **8**, 1–4 (2022).

30. Galanis, M., Devenish, R. J. & Nagley, P. Duplication of leader sequence for protein targeting to mitochondria leads to increased import efficiency. *FEBS Lett.* **282**, 425–430 (1991).

31. Claros, M. G. *et al.* Limitations to in vivo Import of Hydrophobic Proteins into Yeast Mitochondria. *Eur. J. Biochem.* **228**, 762–771 (1995).

32. Wilcox, A. J., Choy, J., Bustamante, C. & Matouschek, A. Effect of protein structure on mitochondrial import. *Proc. Natl. Acad. Sci.* **102**, 15435–15440 (2005).

33. Chin, R. M., Panavas, T., Brown, J. M. & Johnson, K. K. Optimized Mitochondrial Targeting of Proteins Encoded by Modified mRNAs Rescues Cells Harboring Mutations in mtATP6. *Cell Rep.* **22**, 2818–2826 (2018).

34. Rizzuto, R., Simpson, A. W. M., Brini, M. & Pozzan, T. Rapid changes of mitochondrial Ca2+ revealed by specifically targeted recombinant aequorin. *Nature* **358**, 325–327 (1992).

35. Hartl, F.-U., Pfanner, N., Nicholson, D. W. & Neupert, W. Mitochondrial protein import. *Biochim. Biophys. Acta BBA - Rev. Biomembr.* **988**, 1–45 (1989).

36. Betin, V. M. S., MacVicar, T. D. B., Parsons, S. F., Anstee, D. J. & Lane, J. D. A cryptic mitochondrial targeting motif in Atg4D links caspase cleavage with mitochondrial import and oxidative stress. *Autophagy* **8**, 664–676 (2012).

37. Csörgő, B. *et al.* A compact Cascade-Cas3 system for targeted genome engineering. *Nat. Methods* **17**, 1183–1190 (2020).

38. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

39. DiCarlo, J. E. *et al.* Yeast Oligo-Mediated Genome Engineering (YOGE). *ACS Synth. Biol.* **2**, 741–749 (2013).

40. Barbieri, E. M., Muir, P., Akhuetie-Oni, B. O., Yellman, C. M. & Isaacs, F. J. Precise Editing at DNA Replication Forks Enables Multiplex Genome Engineering in Eukaryotes. *Cell* **171**, 1453-1467.e13 (2017).

41. Wannier, T. M. *et al.* Improved bacterial recombineering by parallelized protein discovery. *Proc. Natl. Acad. Sci.* **117**, 13689–13698 (2020).

42. Filsinger, G. T. *et al.* Characterizing the portability of phage-encoded homologous recombination proteins. *Nat. Chem. Biol.* **17**, 394–402 (2021).

43. Kong, X. *et al.* Precise genome editing without exogenous donor DNA via retron editing system in human cells. *Protein Cell* **12**, 899–902 (2021).

44. Schubert, M. G. *et al.* High-throughput functional variant screens via in vivo production of single-stranded DNA. *Proc. Natl. Acad. Sci.* **118**, e2018181118 (2021).

45. Zhao, B., Chen, S.-A. A., Lee, J. & Fraser, H. B. Bacterial Retrons Enable Precise Gene Editing in Human Cells. *CRISPR J.* **5**, 31–39 (2022).

46. Lopez, S. C., Crawford, K. D., Lear, S. K., Bhattarai-Kline, S. & Shipman, S. L. Precise genome editing across kingdoms of life using retron-derived DNA. *Nat. Chem. Biol.* **18**, 199–206 (2022).

47. Rhee, H.-W. *et al.* Proteomic Mapping of Mitochondria in Living Cells via Spatially Restricted Enzymatic Tagging. *Science* **339**, 1328–1331 (2013).

48. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nat. Methods* **18**, 100–106 (2021).

49. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nat. Methods* **19**, 1634–1641 (2022).

50. Dunn, K. W., Kamocka, M. M. & McDonald, J. H. A practical guide to evaluating colocalization in biological microscopy. *Am. J. Physiol. Cell Physiol.* **300**, C723-742 (2011).

51. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, 2009).

52. Kluyver, T. *et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. in (eds. Loizides, F. & Scmidt, B.) 87–90 (IOS Press, 2016). doi:10.3233/978-1-61499-649-1-87.

53. Summerfield, M. Rapid GUI programming with Python and Qt: the definitive guide to PyQt programming. (2007).

Chapter 5: Conclusion

Interest in DNA writers has exponentially increased over the past couple decades, culminating in impressive advances in gene editing, including FDA-approved therapies. Nonetheless, current DNA writing-based tools still have limitations. Our work described here extends the of use of DNA writing tools to broaden its medical applications, specifically by tracking cellular development and increasing the therapeutic use of such tools to treat genetic disease. We first establish how other researchers may accessibly use our molecular recording technology to track the timing of transcriptional events using equipment and expertise typical of most molecular biology and bioinformatics labs. Next, we optimize methods to overcome current obstacles to achieve mitochondrial gene editing, first by increasing HR-directed precise edits using a novel gene editing tool and developing a high-throughput pipeline to screen how efficiently different DNA writers are imported to mammalian mitochondria. This work thus enables other researchers to utilize important tools to mimic cellular development *ex vivo* and lead to more reliable and robust treatments against mitochondrial genetic disease.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Sierra Lear
EA3CE5E3F74A447... Author Signature

7/4/2023
Date